Historical Overview of Audio Spectral Modeling

Julius Smith
CCRMA, Stanford University

Music 421 Applications Lecture

April 2, 2018

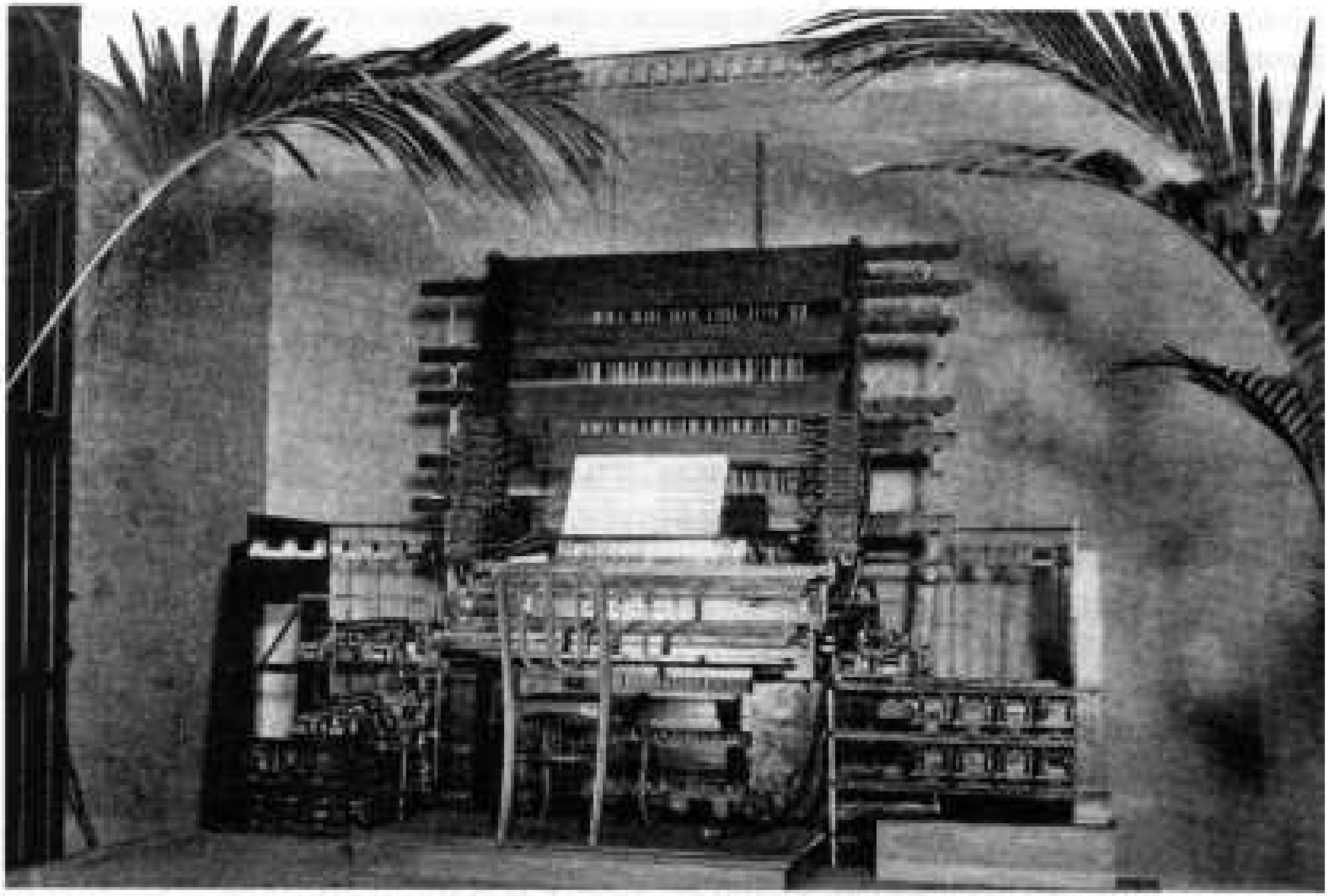
Milestones in Audio Spectral Modeling

- Fourier's theorem (1822)
- Telharmonium (1898)
- Voder (1920s)
- Vocoder (1920s)
- Hammond Organ (1930s)
- Phase Vocoder (1966)
- Digital Organ (1968)
- Additive Synthesis (1969)
- FM Brass Synthesis (1970)
- Synclavier 8-bit FM/Additive synthesizer (1975)
- FM singing voice (1978)
- Sinusoidal Modeling (1985)
- Sines + Noise (1988)
- Sines + Noise + Transients (1988,1996,1998,2000)
- Inverse FFT synthesis (1992)
- Spectrogram Synthesis (2017)
- Future Directions

Telharmonium (1898)

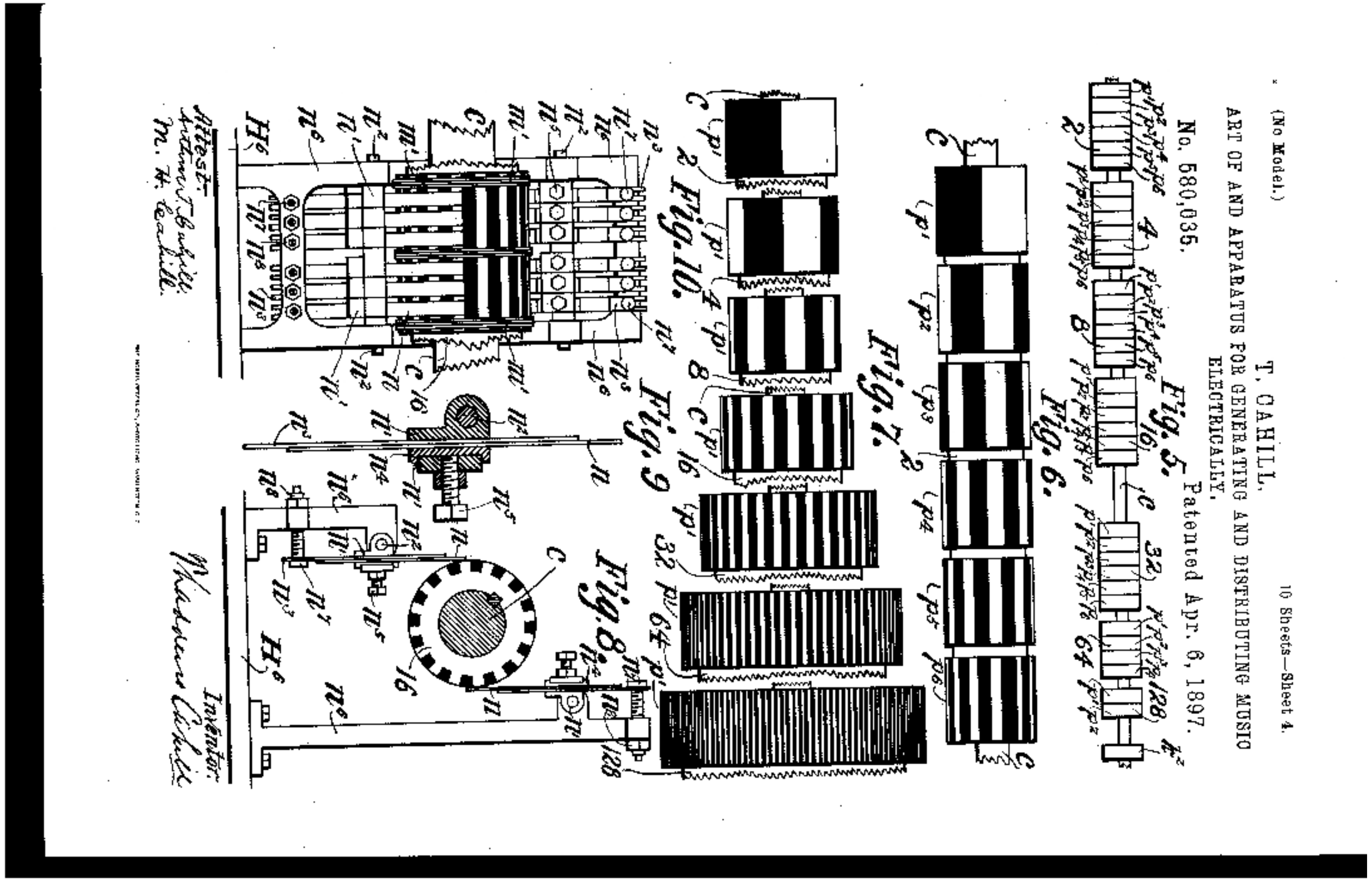
Telharmonium (Cahill 1898)

U.S. patent 580,035:
"Art of and Apparatus for Generating and Distributing Music Electrically"

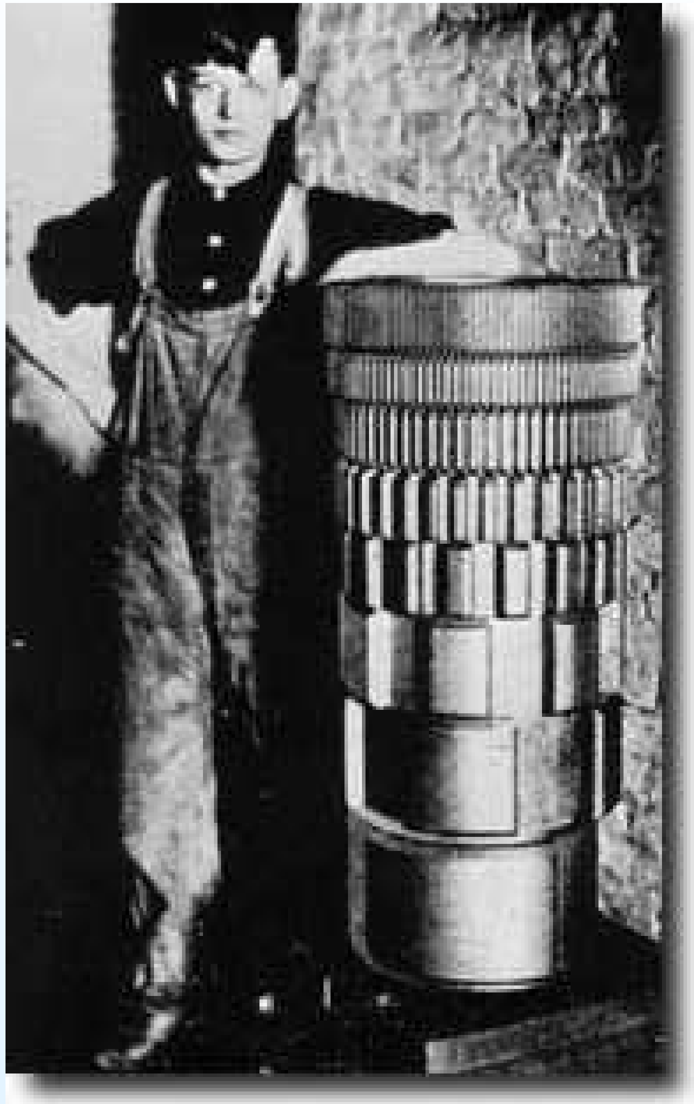


Telharmonium Rheotomes

Forerunner of the Hammond Organ Tone Wheels



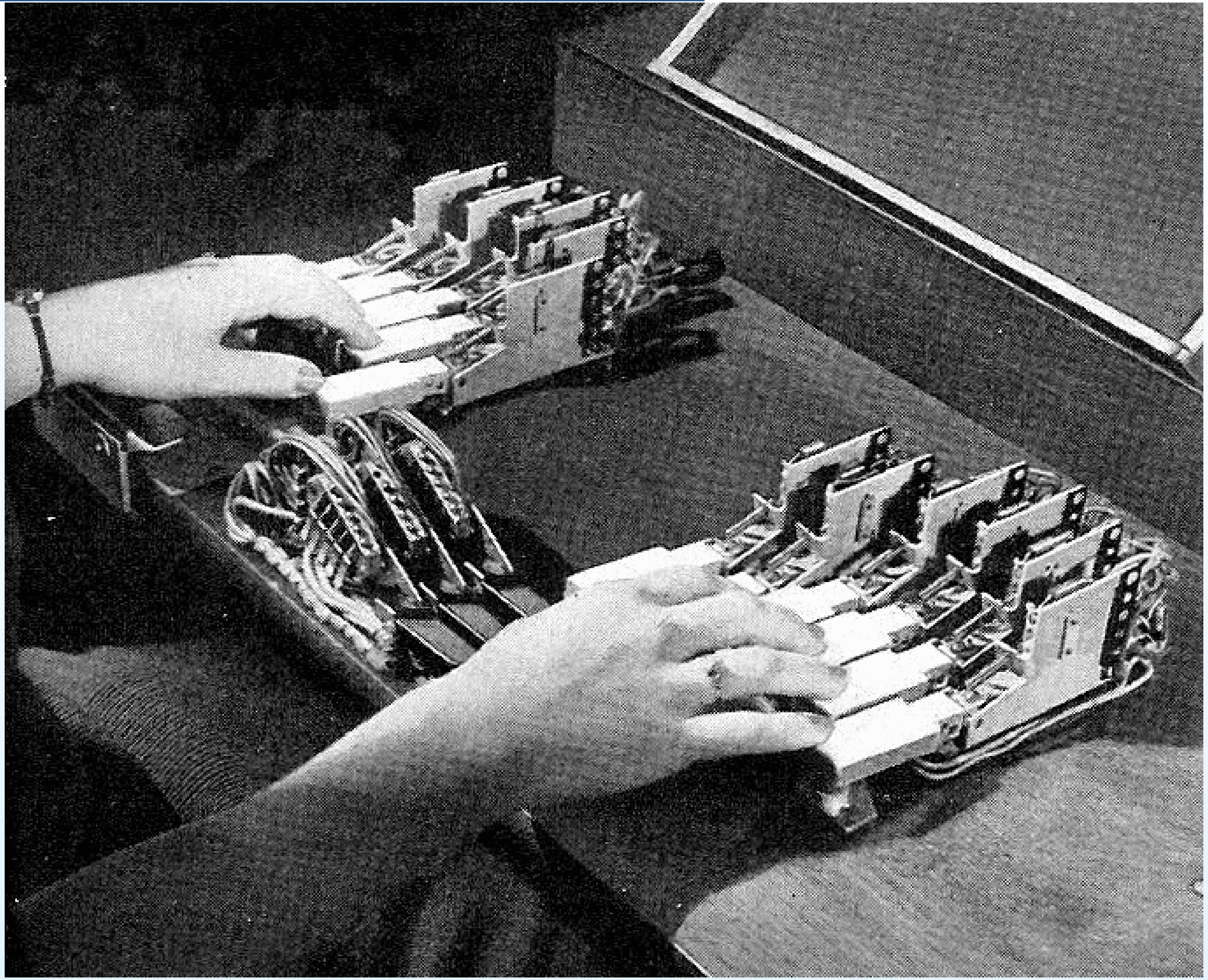
Telharmonium Rotor (early “Tonewheel”)



Hammond influenced: `https://en.wikipedia.org/wiki/Tonewheel`

The Voder (1939)

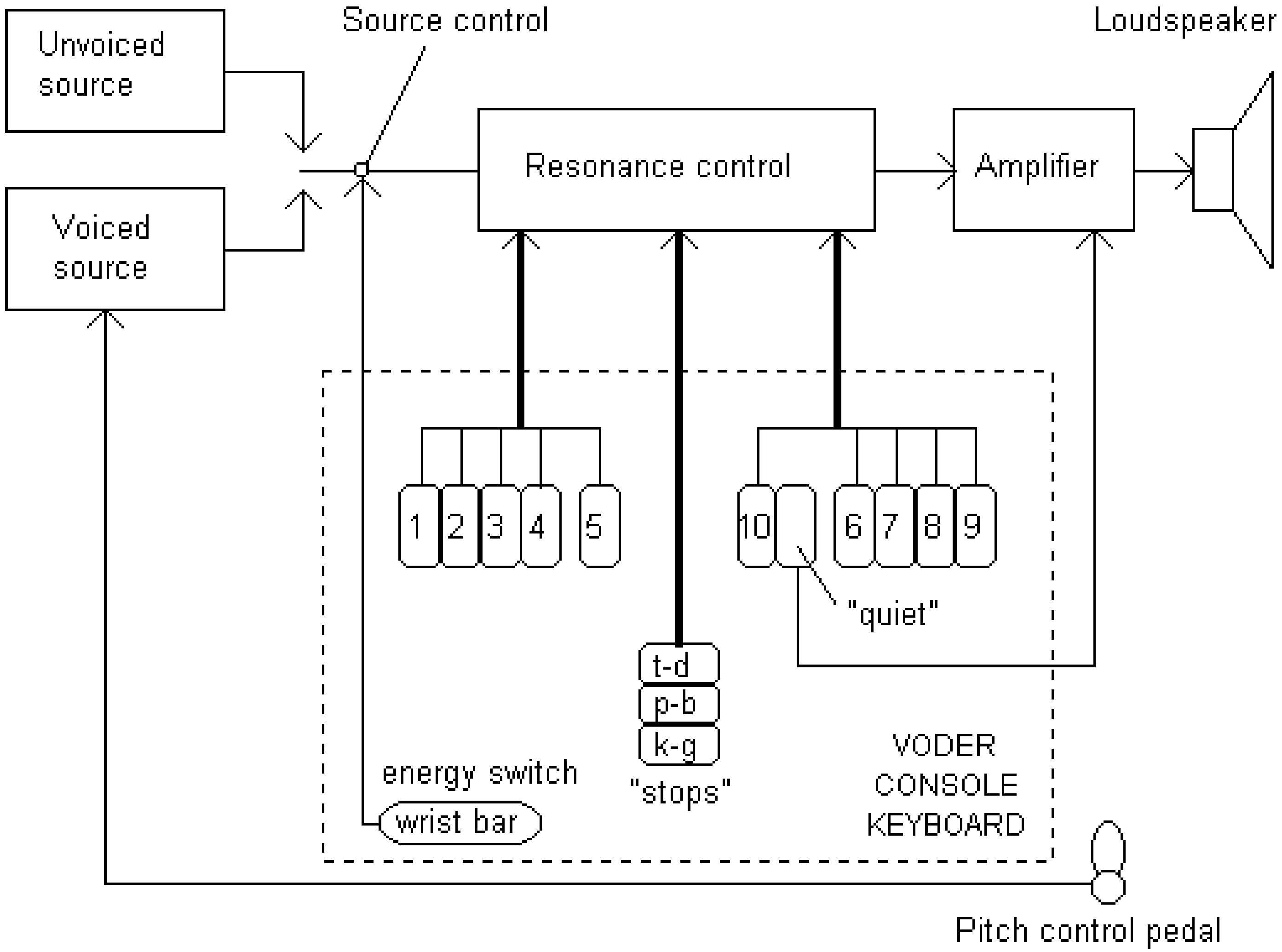
The Voder (Homer Dudley — 1939 Worlds Fair)



$>$

http://davidszondy.com/future/robot/voder.htm

Voder Keyboard



http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/chap2.html — (from Klatt 1987)

Voder Schematic

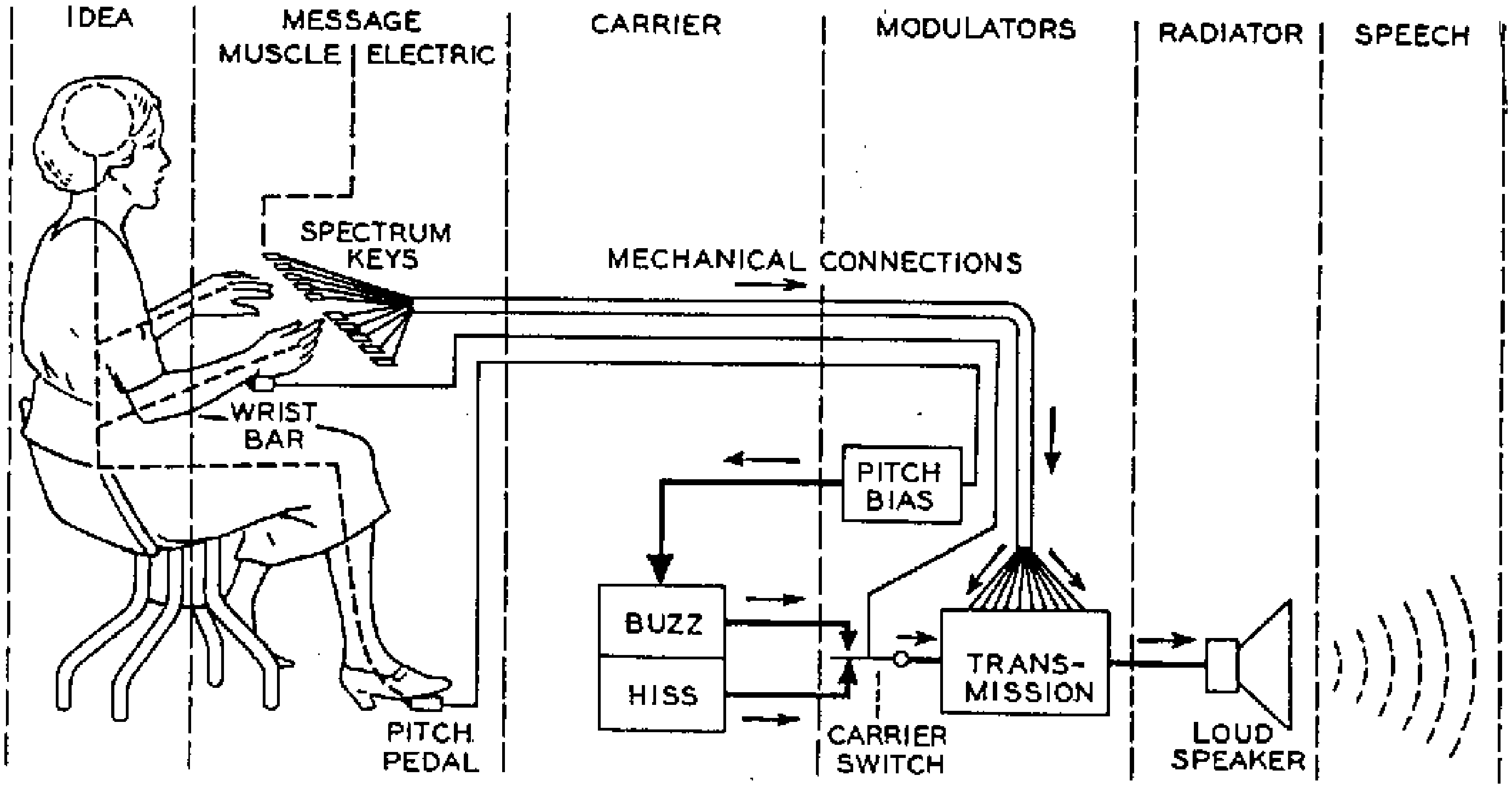


Fig. 8—Schematic circuit of the voder.

http://ptolemy.eecs.berkeley.edu/~eal/audio/voder.html

Voder Demos

- Voder Demo (Audio and Video)
- More Voder Demos - Audio Only [Demos Begin]

The Channel Vocoder (1928)
(“Voice Coder”)

Vocoder Analysis & Resynthesis (Dudley 1928)

Analysis:

- Ten analog bandpass filters between 250 and 3000 Hz: Bandpass $\rightarrow$ rectifier $\rightarrow$ lowpass filter $\rightarrow$ *amplitude envelope*
- Voiced/Unvoiced decision made
- Fundamental frequency F_0 measured for voiced case

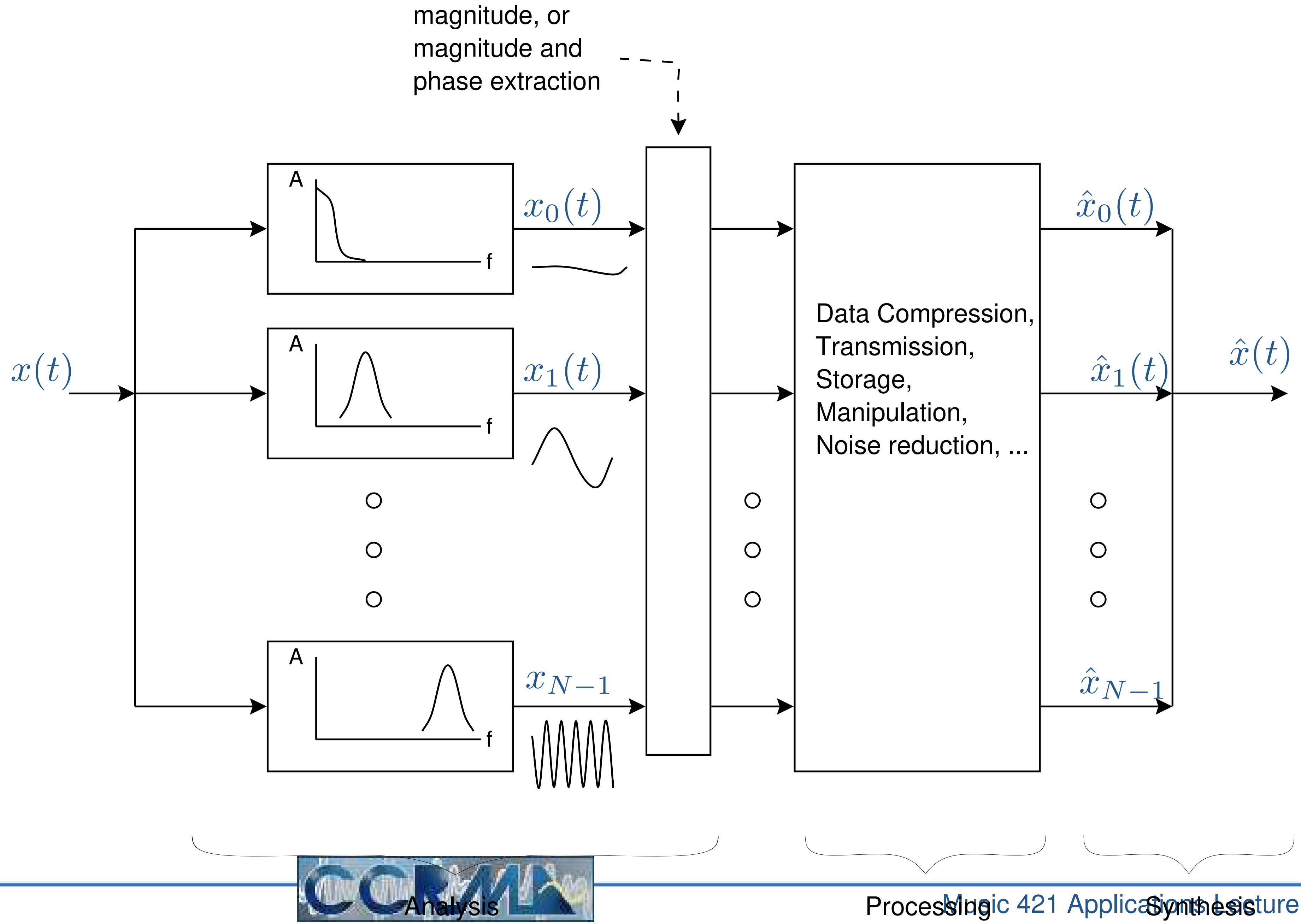
Synthesis:

- Ten matching bandpass filters driven by a
  - "buzz source" (voiced), or
  - "hiss source" (unvoiced)
- Bands were scaled by amplitude envelopes and summed
- Said to have an "unpleasant electrical accent"

Related Speech Models:

- The Vocoder is an early *source-filter* model for speech
- *Linear Predictive Coding* (LPC) of speech is another

Vocoder Filter Bank Analysis/Resynthesis



Channel Vocoder Sound Examples

- Original

- 10 channels, sine carriers
- 10 channels, narrowband-noise carriers

- 26 channels, sine carriers
- 26 channels, narrowband-noise carriers
- 26 channels, narrowband-noise carriers, channels reversed

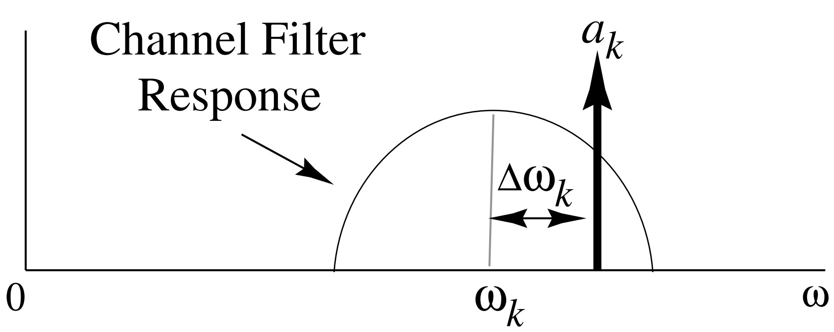
- ***Phase* Vocoder:** Identity system in absence of modifications

- The FFT Phase Vocoder next transitioned to the Short-Time Fourier Transform (STFT) (Allen and Rabiner 1977)

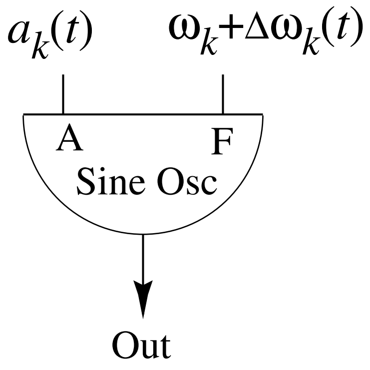
The Phase Vocoder (1966)

Phase Vocoder Analysis for Additive Synthesis (1976)

Analysis Model

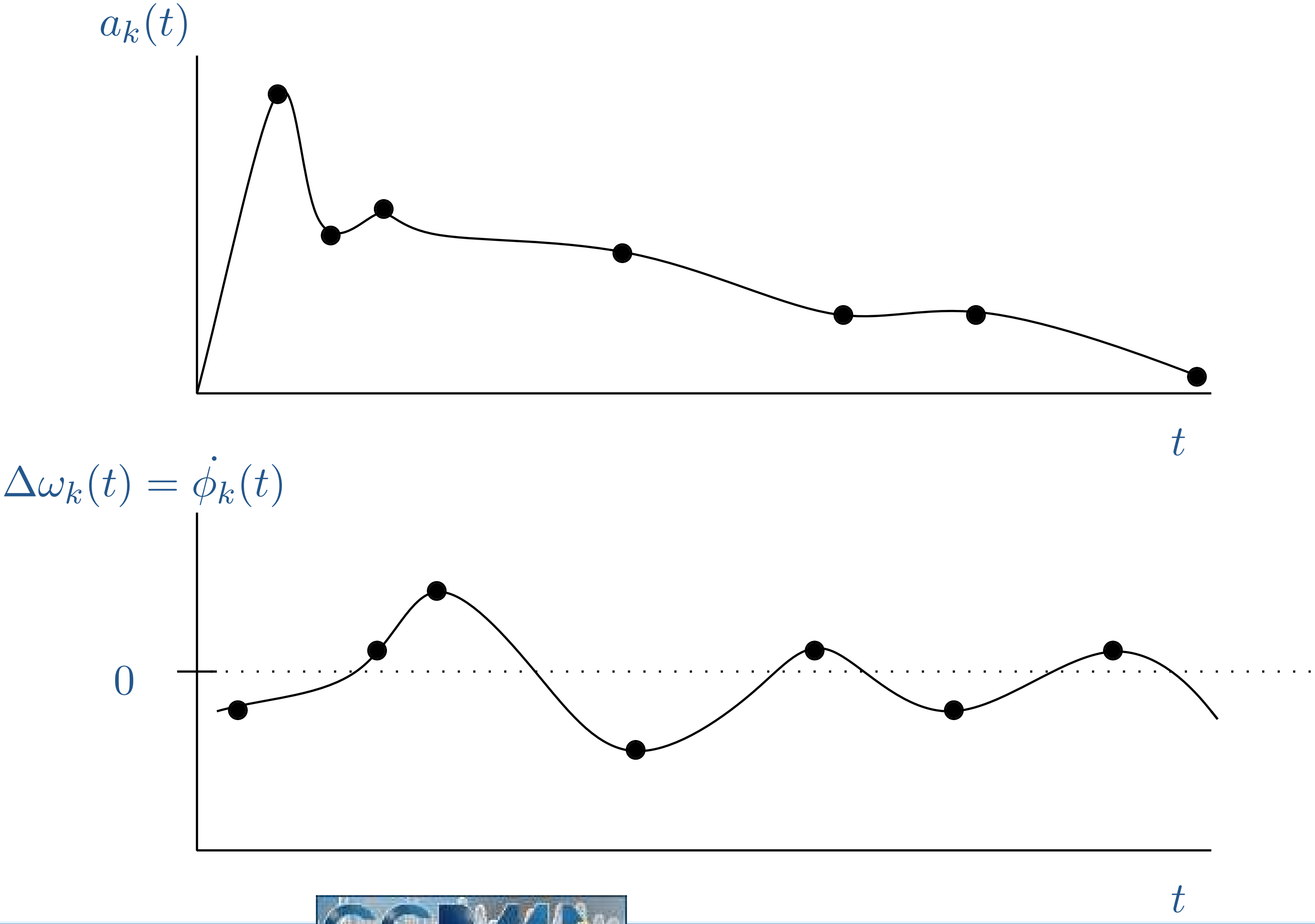


Synthesis Model



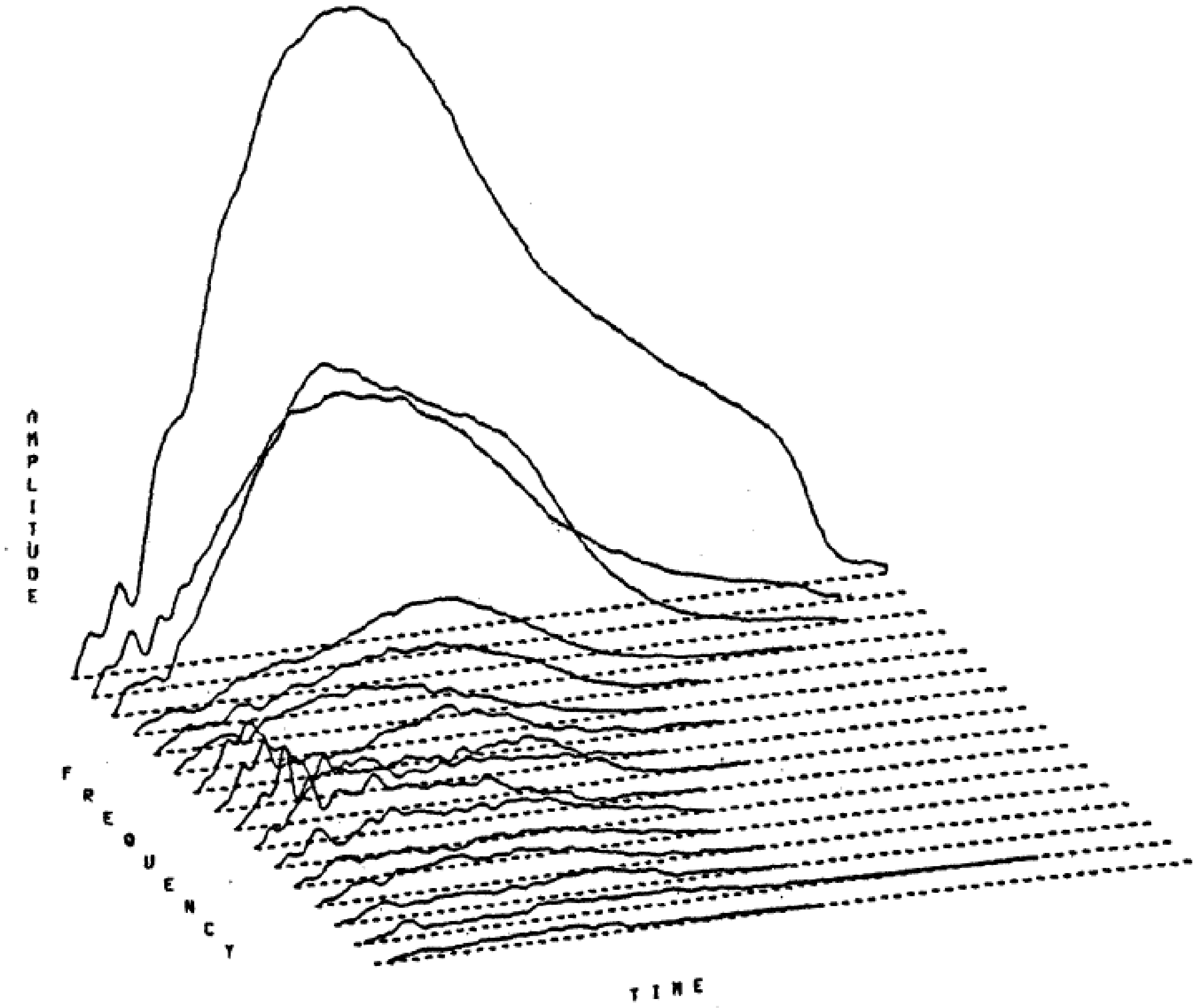
- Early "channel vocoder" implementations (hardware) only measured amplitude $a_k(t)$ (Dudley 1939)
- The "phase vocoder" (Flanagan and Golden 1966) added phase tracking in each channel
- Portnoff (1976) developed the FFT phase vocoder, replacing the heterodyne comb in computer-music additive-synthesis analysis (James A. Moorer 1975)
- Inverse FFT synthesis (Rodet and Depalle 1992) gave faster sinusoidal oscillator banks

Amplitude and Frequency Envelopes



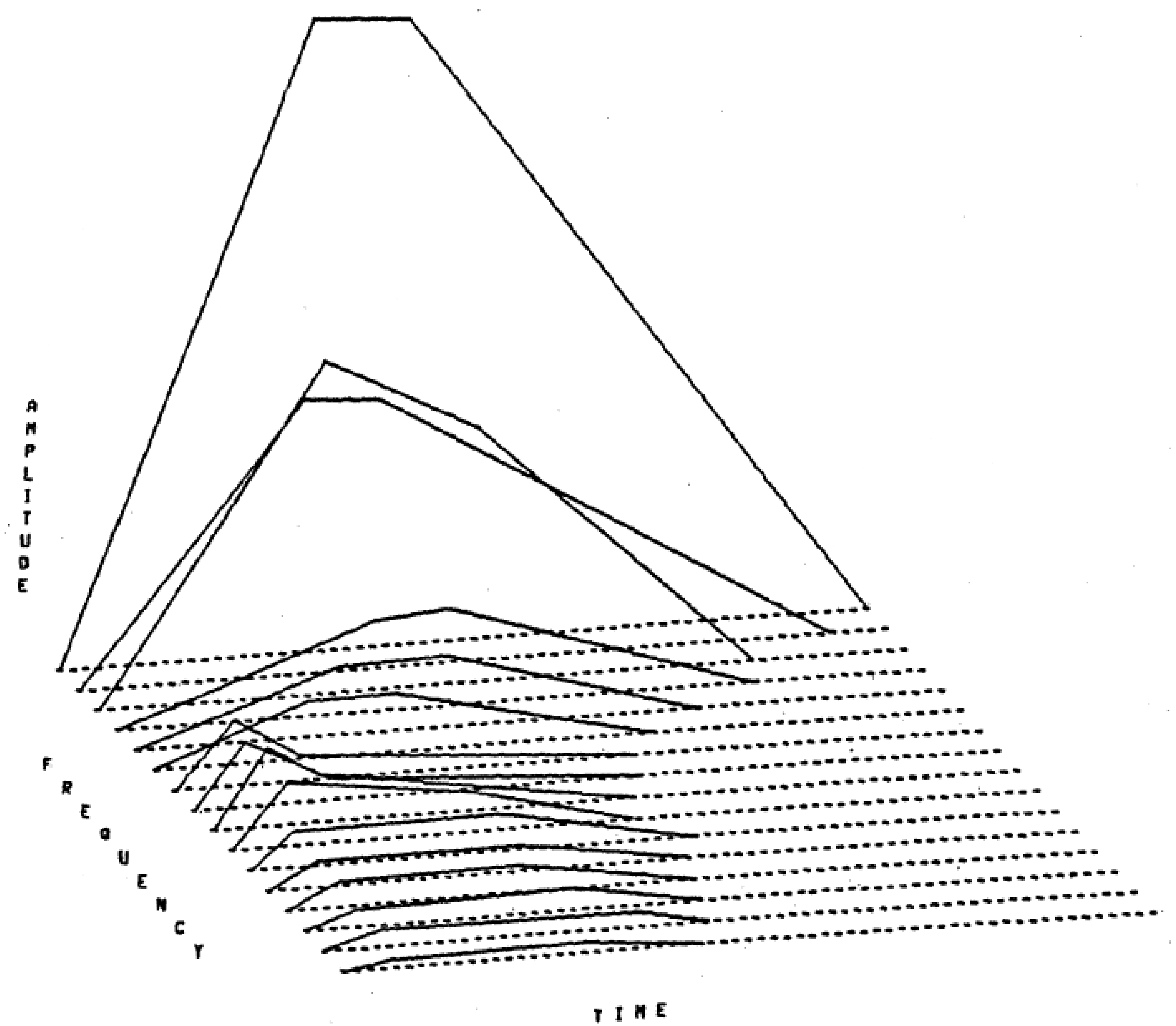
Additive Synthesis (1969)

Classic Additive-Synthesis Analysis (Heterodyne Comb)



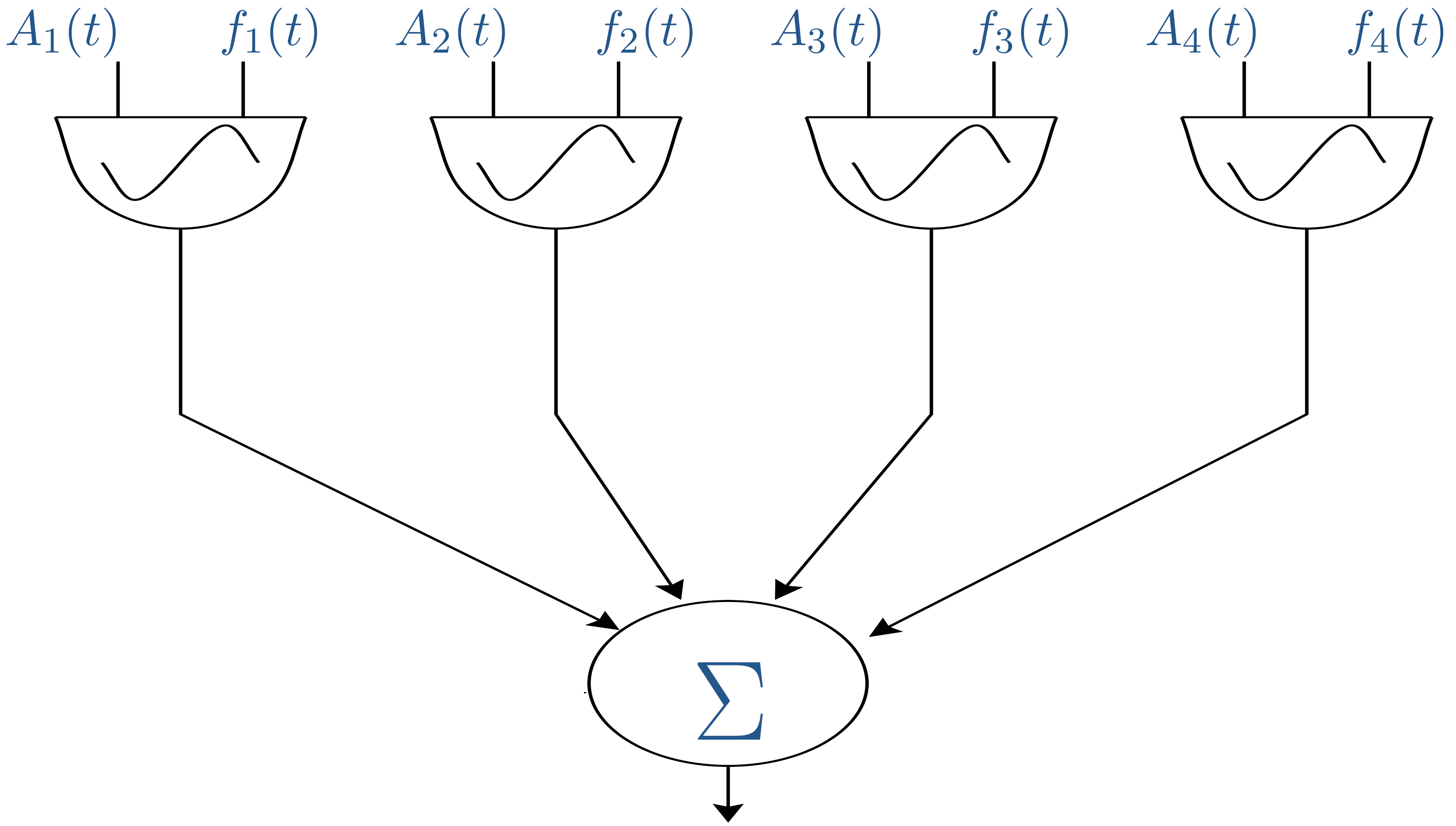
John Grey 1975 — CCRMA Tech. Reports 1 & 2
(CCRMA "STANM" reports — available online)

Classic Additive-Synthesis (Sinusoidal Oscillator Envelopes)



John Grey 1975 — CCRMA Tech. Reports 1 & 2
(CCRMA “STANM” reports available online)

Classic Additive Synthesis Diagram (Computer Music, 1960s)



$$y(t) = \sum_{i=1}^{4} A_i(t) \sin\left[\int_0^t \omega_i(t)dt + \phi_i(0)\right]$$

Classic Additive-Synthesis Examples

- Bb Clarinet
- Eb Clarinet
- Oboe
- Bassoon
- Tenor Saxophone
- Trumpet
- English Horn
- French Horn
- Flute

- All of the above
- Independently synthesized set

(Synthesized from original John Grey data)

Frequency Modulation Synthesis (1973)

Frequency Modulation (FM) Synthesis

FM synthesis is normally used as a *spectral modeling* technique

- Discovered and developed (1970s) by John M. Chowning (CCRMA Founding Director)
- Key paper: JAES 1973 (vol. 21, no. 7)
- Commercialized by Yamaha Corporation:
  - DX-7 synthesizer (1983)
  - OPL chipset (SoundBlaster PC sound card)
  - Cell phone ring tones

- On the physical modeling front, synthesis of vibrating-string waveforms using *finite differences* started around this time:
  Hiller & Ruiz, JAES 1971 (vol. 19, no. 6)

FM Formula

$$x(t) = A_c \sin[\omega_c t + \phi_c + A_m \sin(\omega_m t + \phi_m)]$$

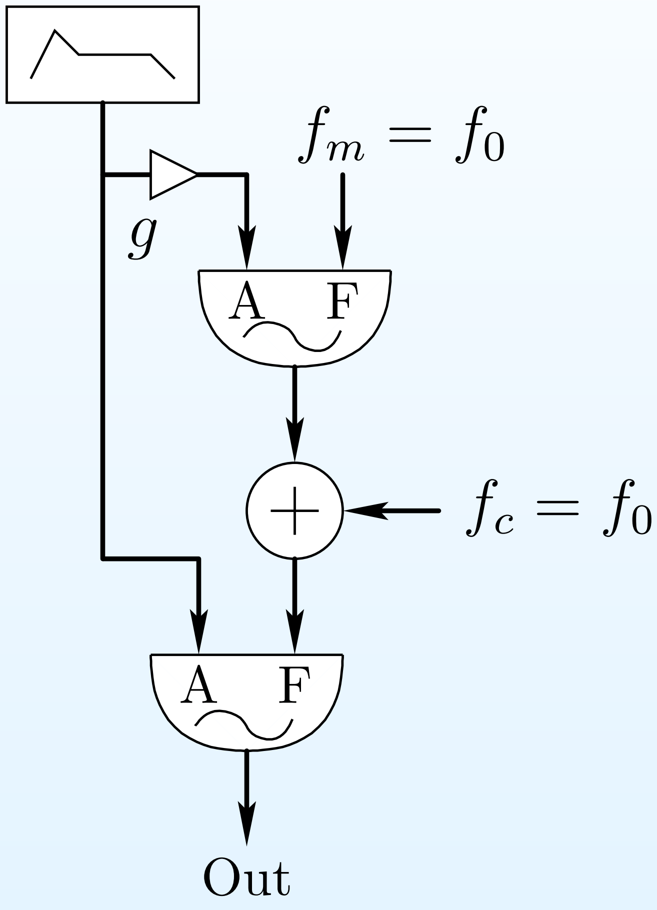
where

(A_c, ω_c, ϕ_c) specify the *carrier* sinusoid
(A_m, ω_m, ϕ_m) specify the *modulator* sinusoid

Can also be called *phase modulation*

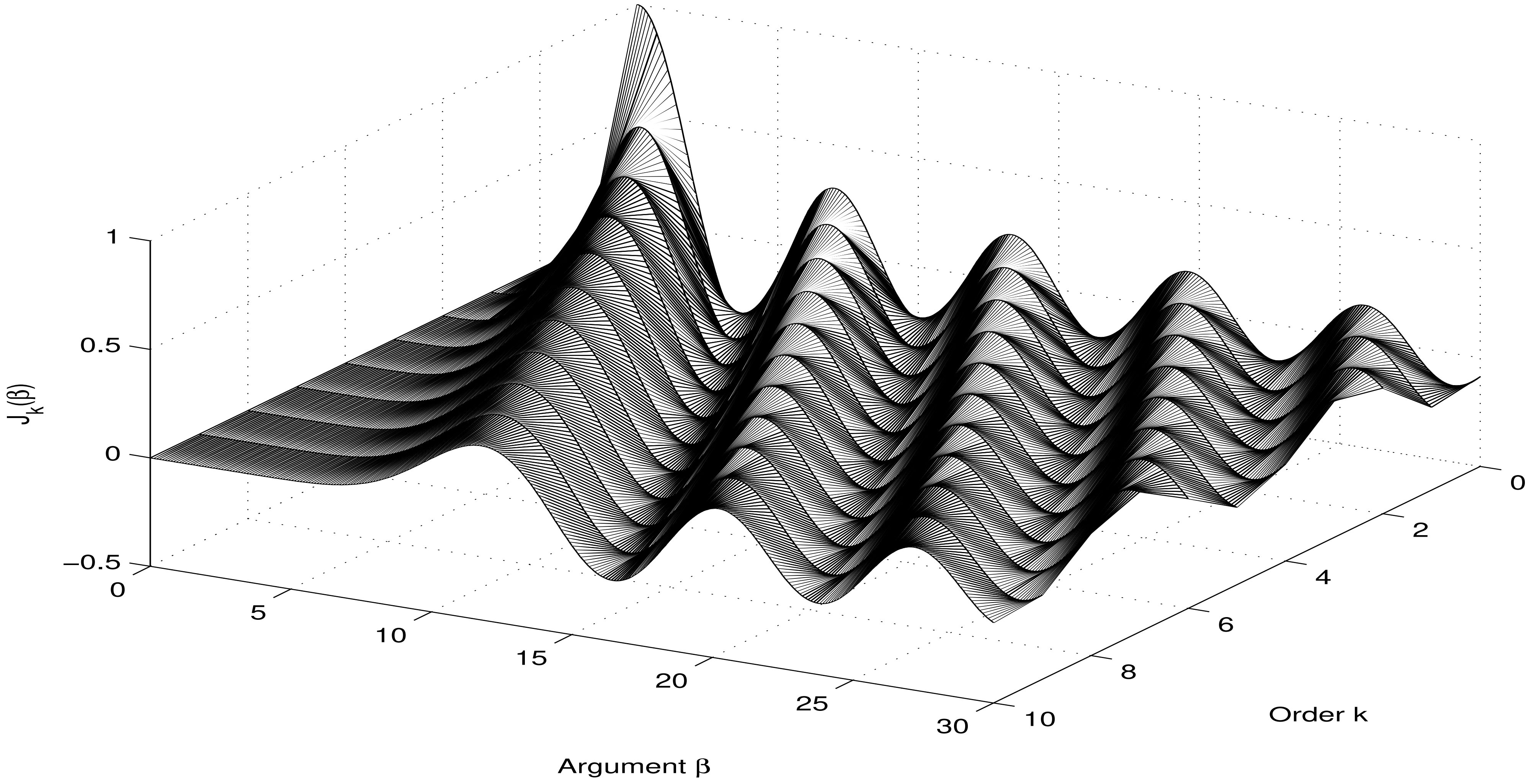
Simple FM "Brass" Patch (Chowning 1970–)

Jean-Claude Risset observation (1964–1969):
Brass bandwidth $\propto$ amplitude



FM Harmonic Amplitudes (Bessel Function of First Kind)

Harmonic number k, FM index β:



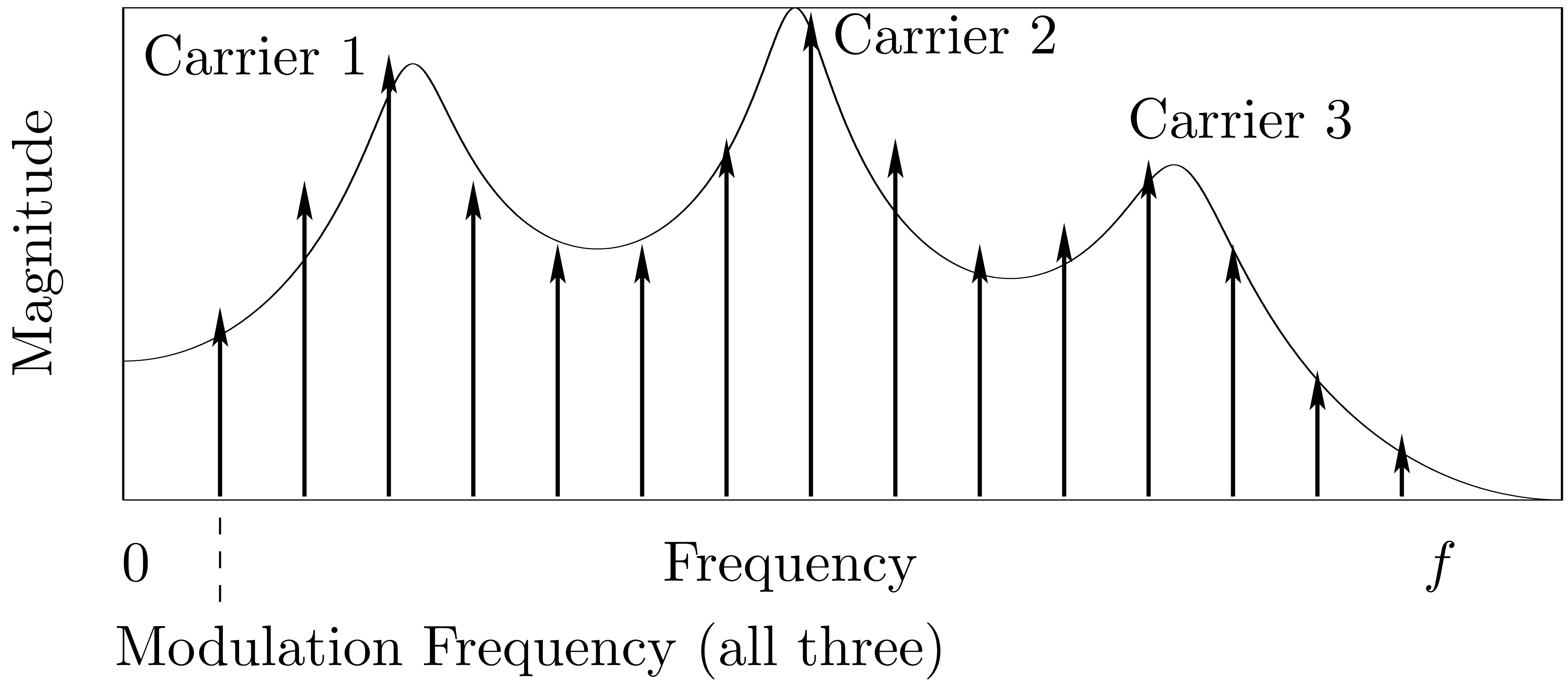
Frequency Modulation (FM) Examples

All examples by John Chowning unless otherwise noted:

- FM brass synthesis
  - Low Brass example
  - Dexter Morril's FM Trumpet
- FM singing voice (1978)
  Each formant synthesized using an FM operator pair
  (two sinusoidal oscillators)
  - Chorus
  - Voices
  - Basso Profundo
- Other early FM synthesis
  - Clicks and Drums
  - Big Bell
  - String Canon

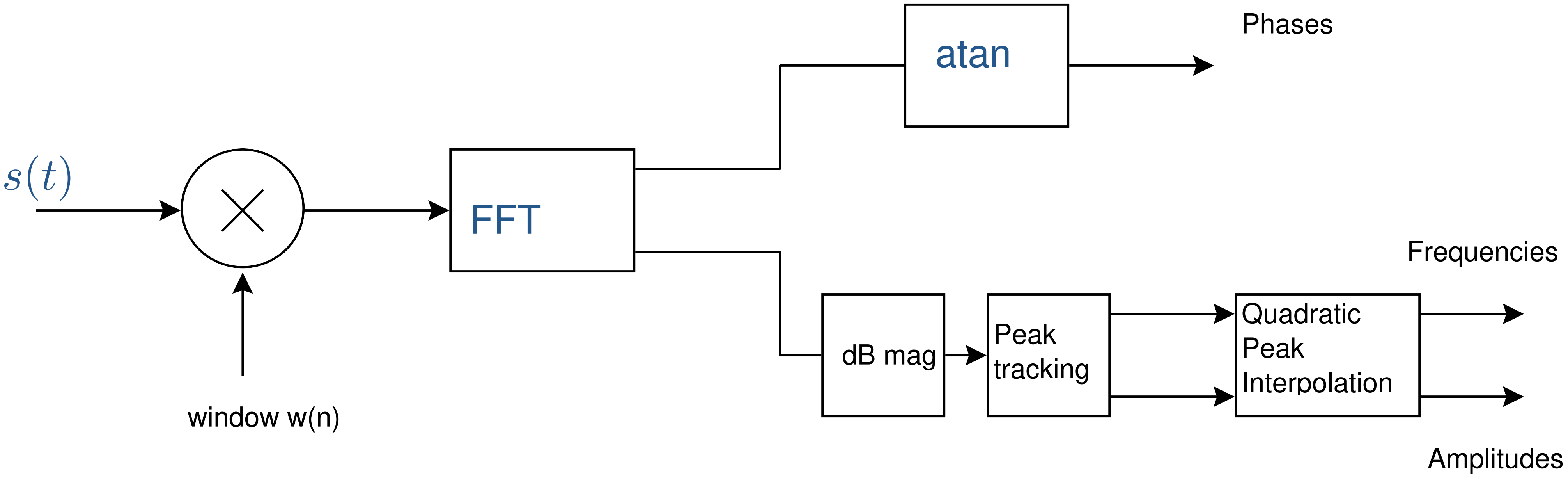
FM Voice

FM voice synthesis can be viewed as *compressed modeling of spectral formants*



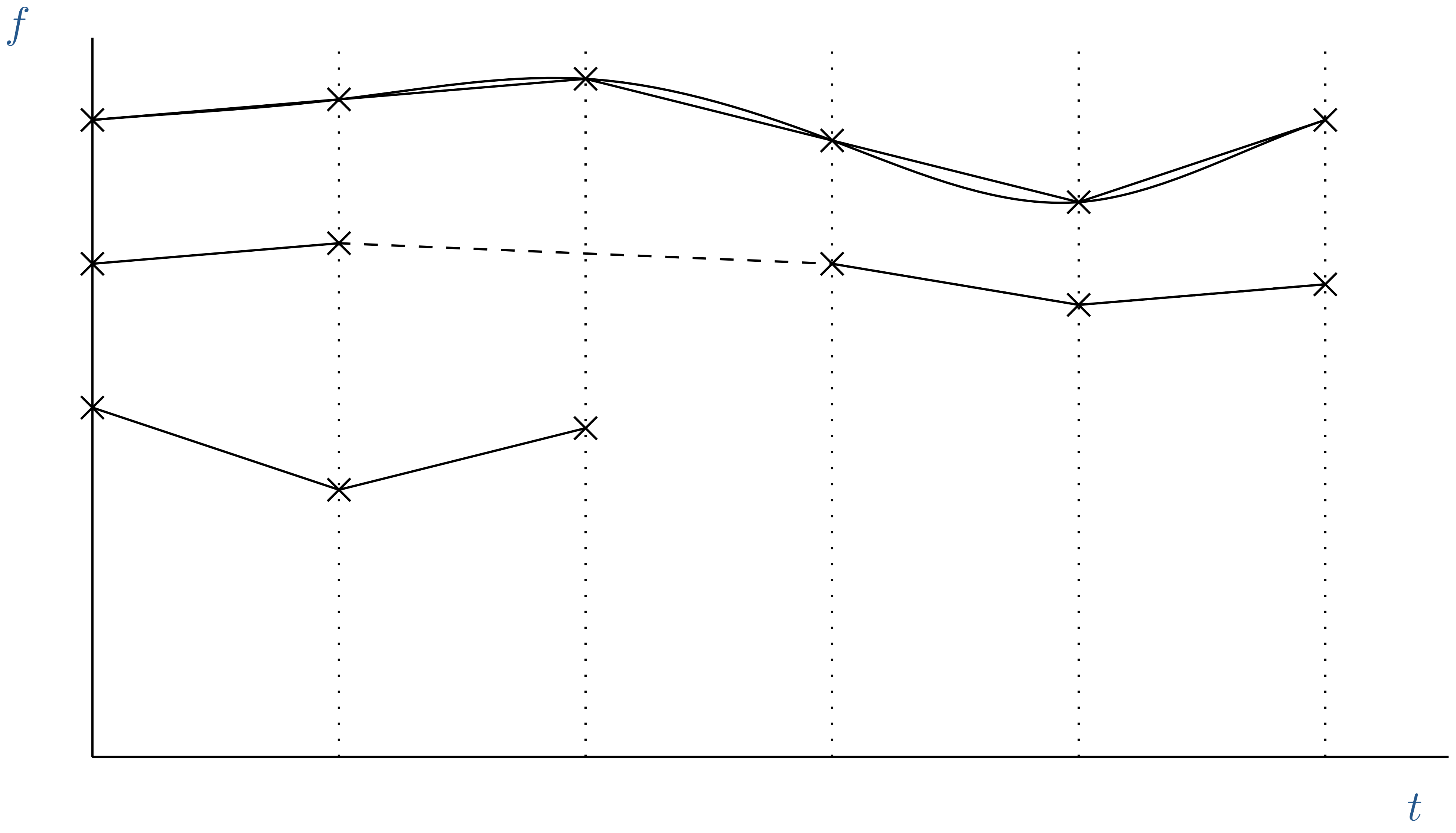
Sinusoidal Modeling Synthesis (1988)

Tracking Spectral Peaks in the Short-Time Fourier Transform

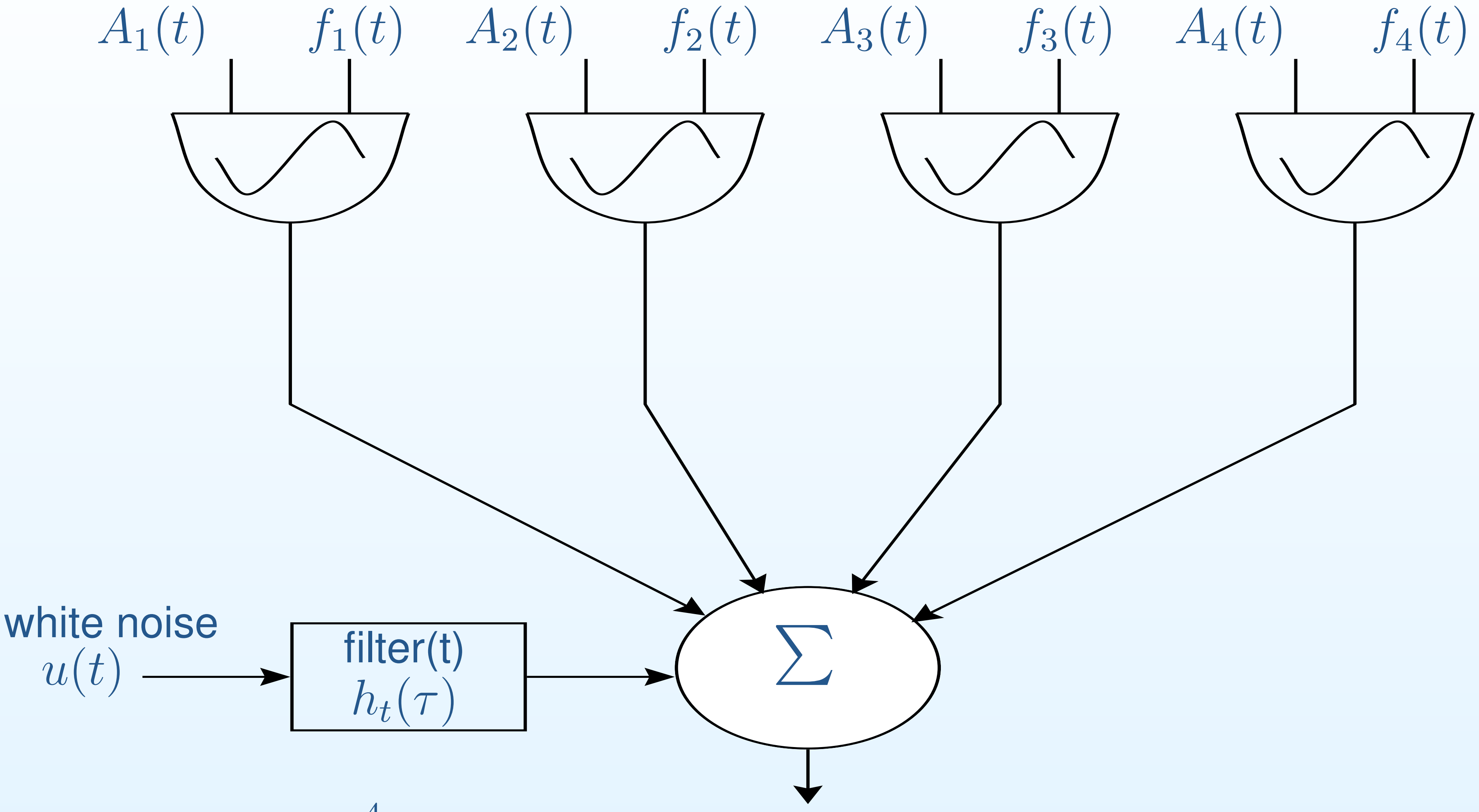


- STFT peak tracking at CCRMA: mid-1980s (PARSHL program)
- Motivated by vocoder analysis of piano tones
- Influences: STFT (Allen and Rabiner 1977), ADEC (1977), MAPLE (1979)
- Independently developed for speech coding by McAulay and Quatieri at Lincoln Labs (1985)

Example Spectral Trajectories



Parametric Spectral Modeling



$$y(t) = \sum_{i=1}^{4} A_i(t) \cos\left[\int_0^t \omega_i(t)dt + \phi_i(0)\right] + (h_t * u)(t)$$

Sines + Noise Sound Examples

Xavier Serra thesis demos (Sines + Noise signal modeling)

- Piano
  - Original
  - Sinusoids alone
  - Residual after sinusoids removed
  - Sines + noise model
- Voice
  - Original
  - Sinusoids
  - Residual
  - Synthesis

Musical Effects with Sines+Noise Models

- Piano Effects
  - Pitch downshift one octave
  - Pitch flattened
  - Varying partial stretching
- Voice Effects
  - Frequency-scale by 0.6
  - Frequency-scale by 0.4 and stretch partials
  - Variable time-scaling, deterministic to stochastic

Cross-Synthesis with Sines+Noise Models

- Voice "modulator"

- Creaking ship's mast "carrier"
- Voice-modulated creaking mast
- Same with modified spectral envelopes

Sines + Transients Sound Examples

In this simple technique, the sinusoidal sum is phase-matched at the cross-over point only (with no cross-fade).

- Marimba
  - Original
  - Sinusoidal model
  - Original attack, followed by sinusoidal model
- Piano
  - Original
  - Sinusoidal model
  - Original attack, followed by sinusoidal model

Multiresolution Sines + Noise + Transients

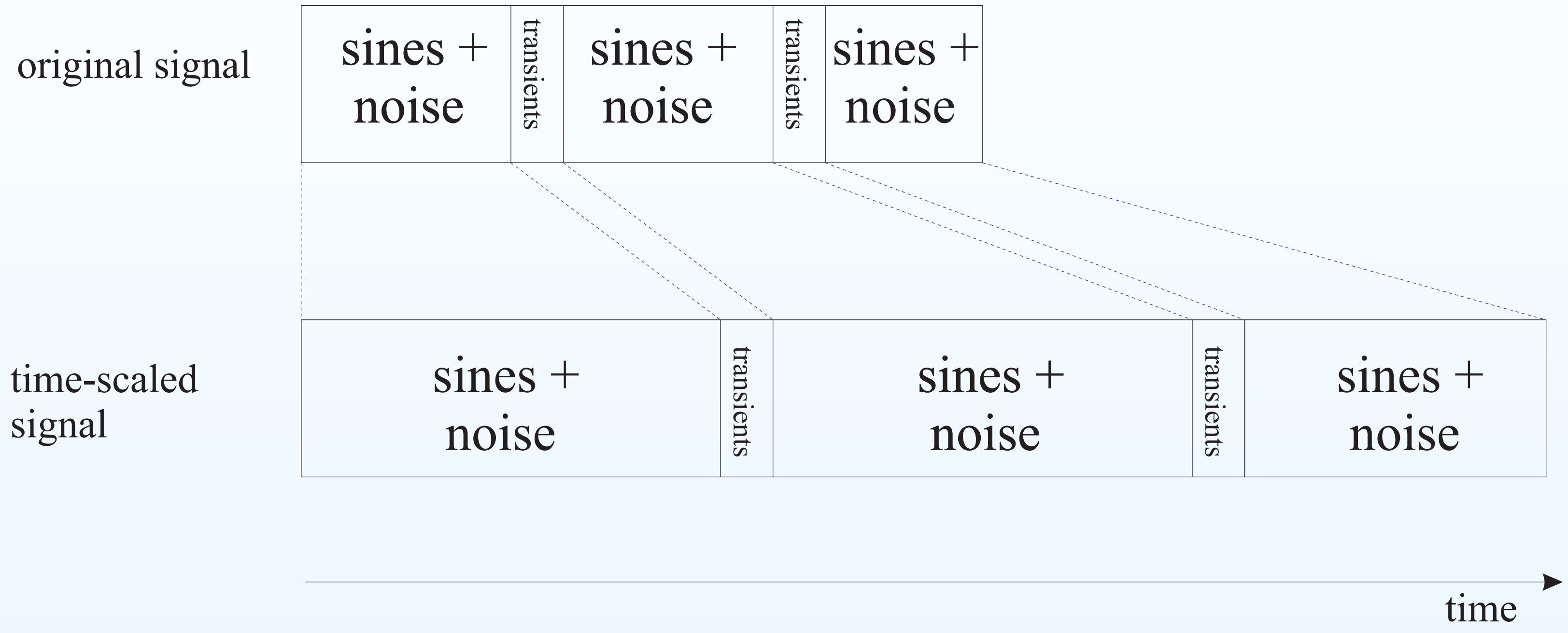
Why Model Transients Separately?

- Sinusoids efficiently model spectral *peaks* over time
- Filtered noise efficiently models spectral *residual* vs. t
- Neither is good for *abrupt transients* in the waveform
- Phase-matched oscillators are expensive
- More efficient to switch to a *transient model* during transients
- Need sinusoidal *phase matching* at the switching times

Transient models:

- Original waveform slice (1988)
- Wavelet expansion (Ali 1996)
- MPEG-2 AAC (with short window) (Levine 1998)
- Frequency-domain LPC
  (time-domain amplitude envelope) (Verma 2000)

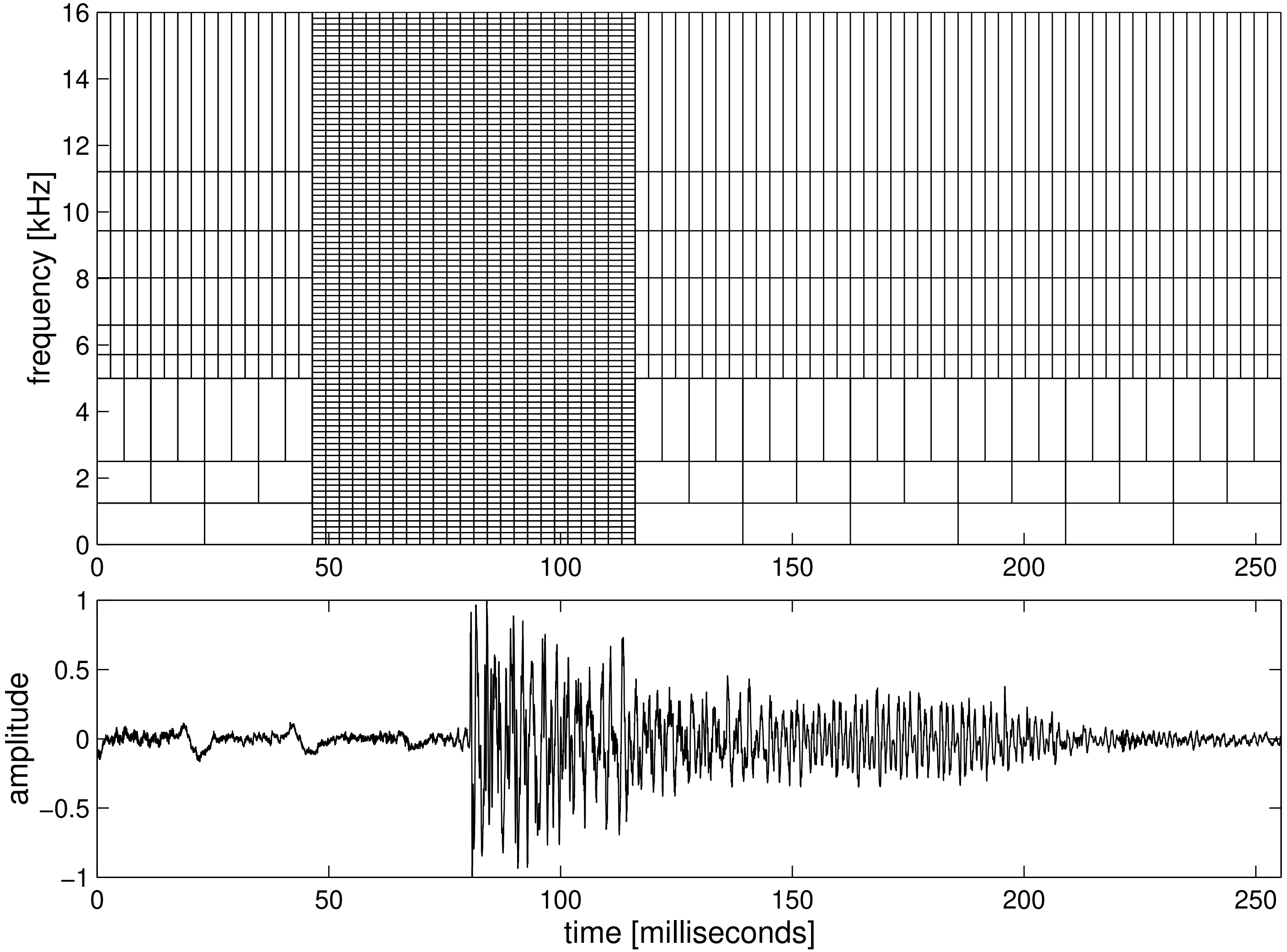
Time Scale Modification of Sines + Noise + Transients Models



Time-Scale Modification (TSM) becomes *well defined*:

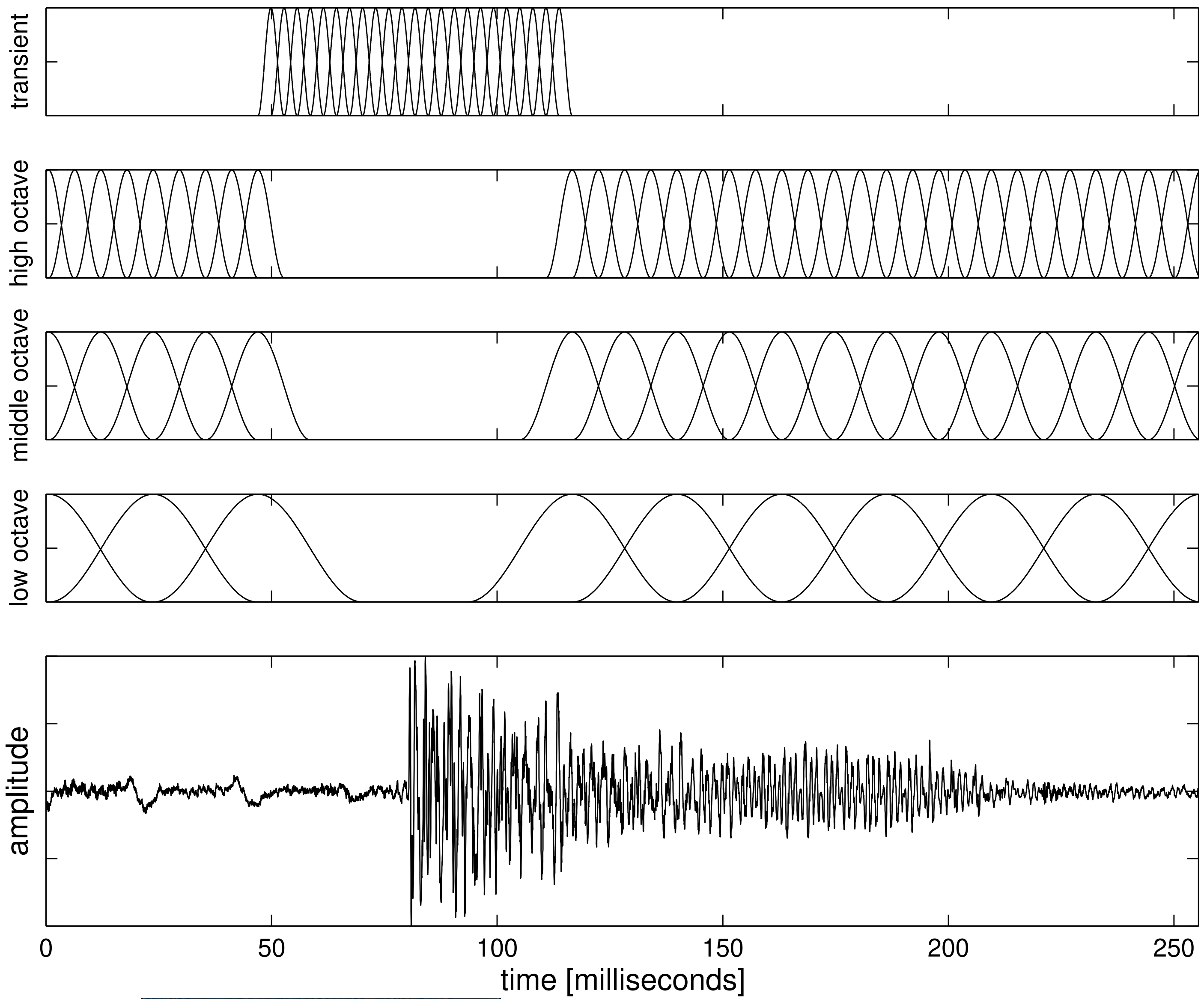
- Transients are *translated* in time
- Sinusoidal envelopes are *scaled* in time
- Noise-filter envelopes also *scaled* in time
- Dual of TSM is *frequency scaling*

Sines + Noise + Transients Time-Frequency Map

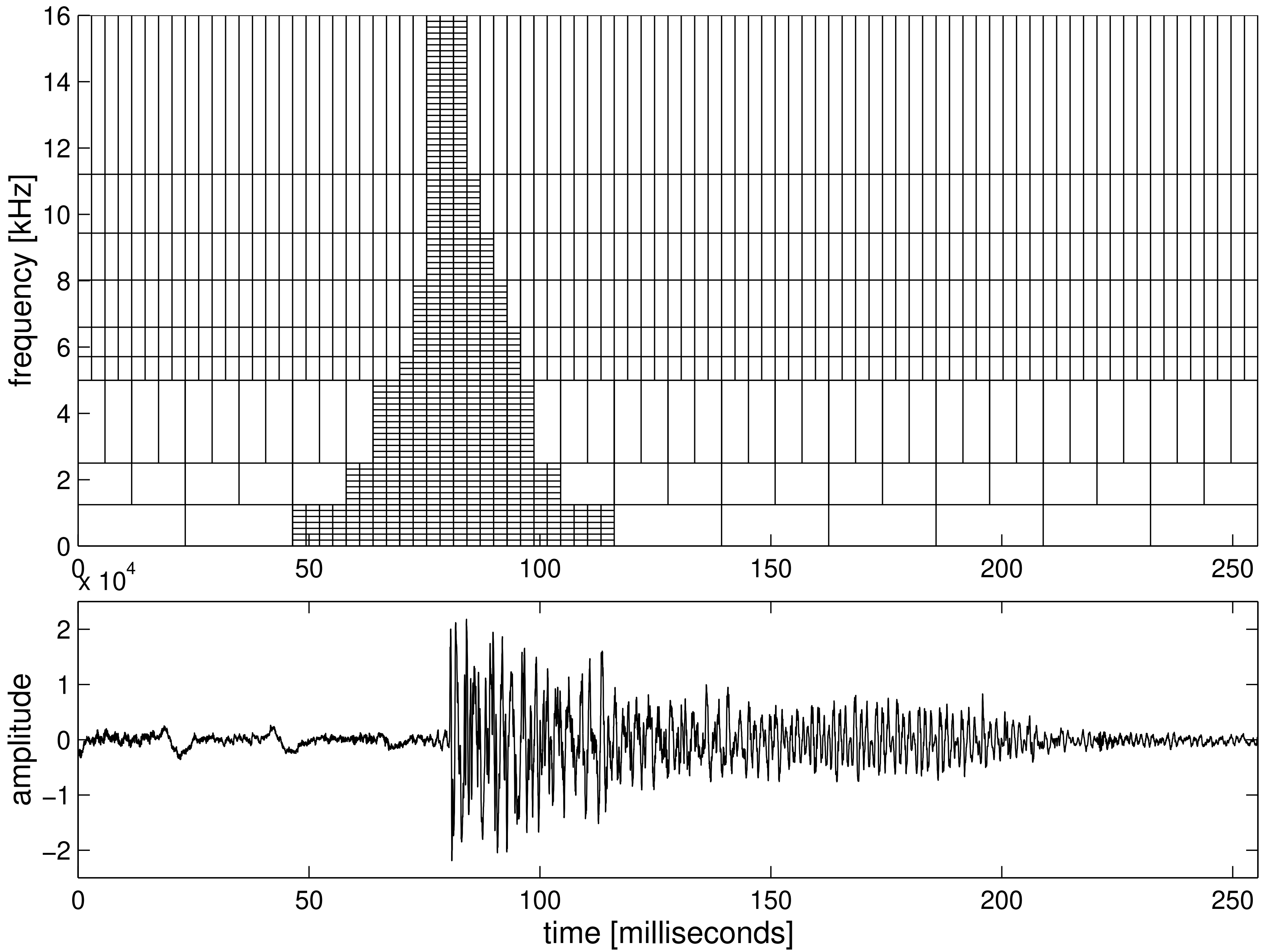


(Levine 1998)

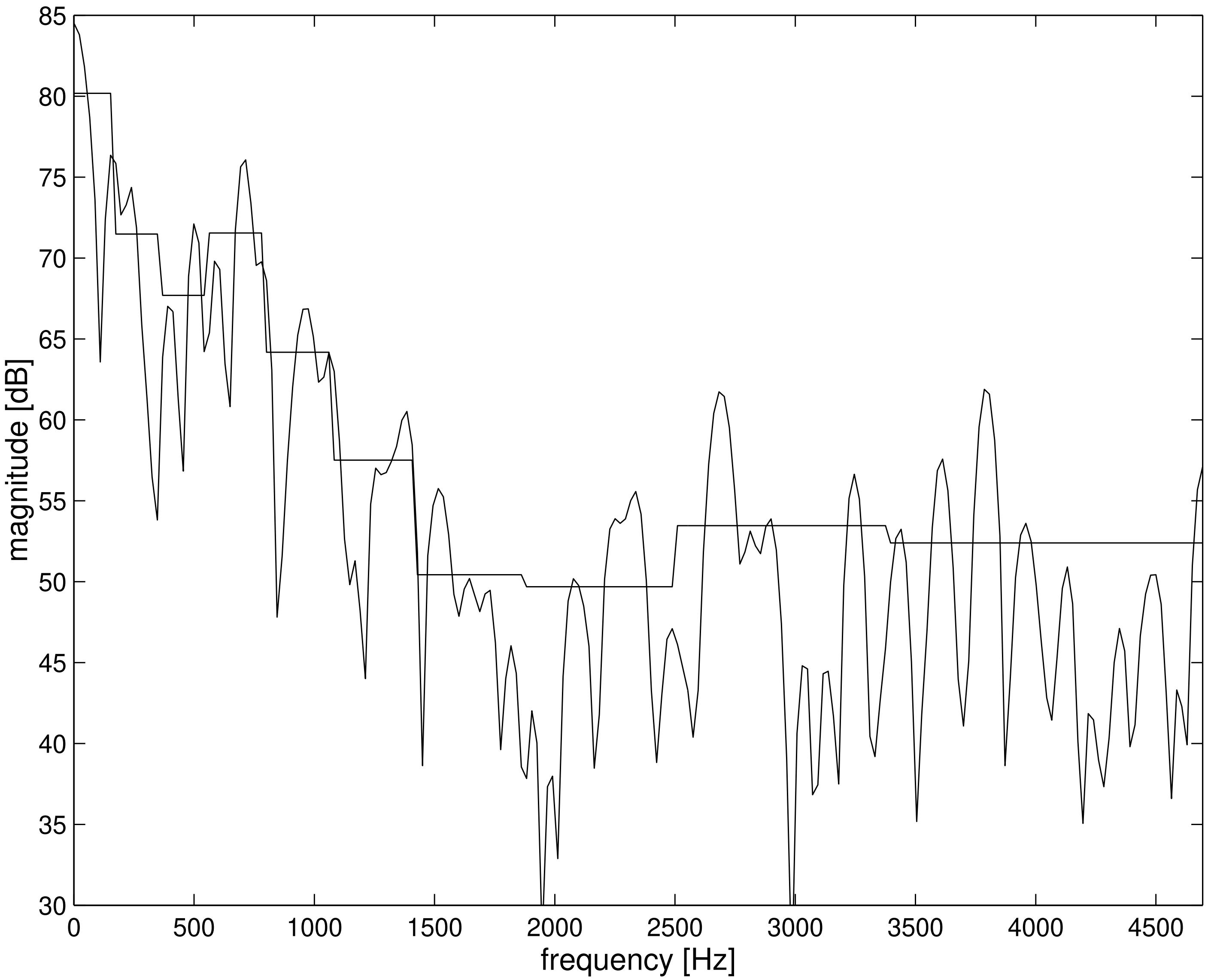
Corresponding Analysis Windows



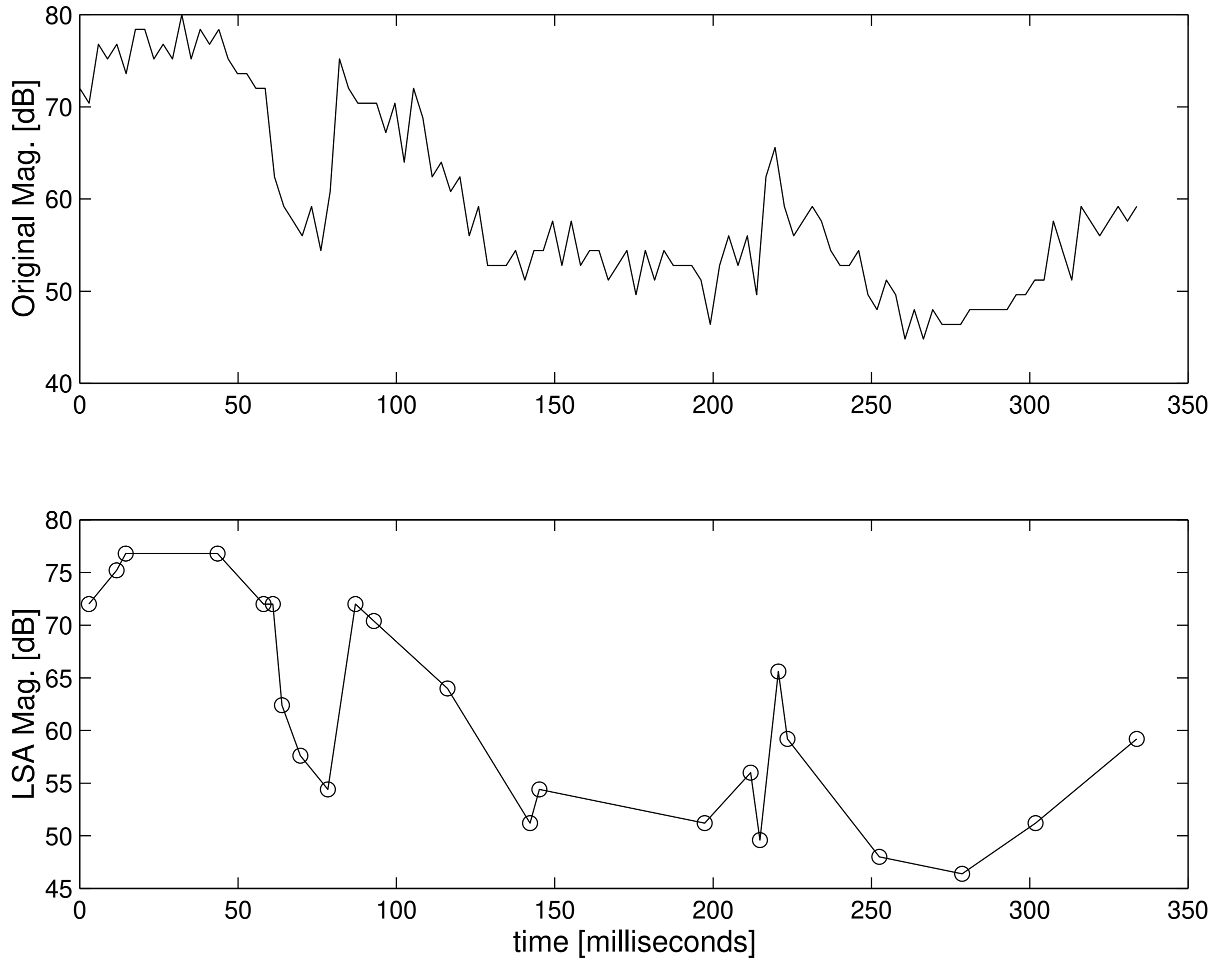
Quasi-Constant-Q (Wavelet) Time-Frequency Map



Bark-Band Noise Modeling (Levine 1998)



Amplitude Envelope for One Noise Band



Sines + Noise + Transients Sound Examples

Scott Levine Thesis Demos (Sines + Noise + Transients at 32 kbps)
(`http://ccrma.stanford.edu/~scottl/thesis.html`)

"It Takes Two" by Rob Base & DJ E-Z Rock

- Original
- MPEG-AAC at 32 kbps
- Sines+transients+noise at 32 kbps

- Multiresolution sinusoids
- Residual Bark-band noise
- Transform-coded transients (AAC)
- Bark-band noise above 5 kHz

Time Scale Modification using Sines + Noise + Transients

Scott Levine Thesis Demos (Sines + Noise + Transients at 32 kbps)
(`http://ccrma.stanford.edu/~scottl/thesis.html`)

Time-Scale Modification (pitch unchanged)

- S+N+T time-scale factors [2.0, 1.6, 1.2, 1.0, 0.8, 0.6, 0.5]

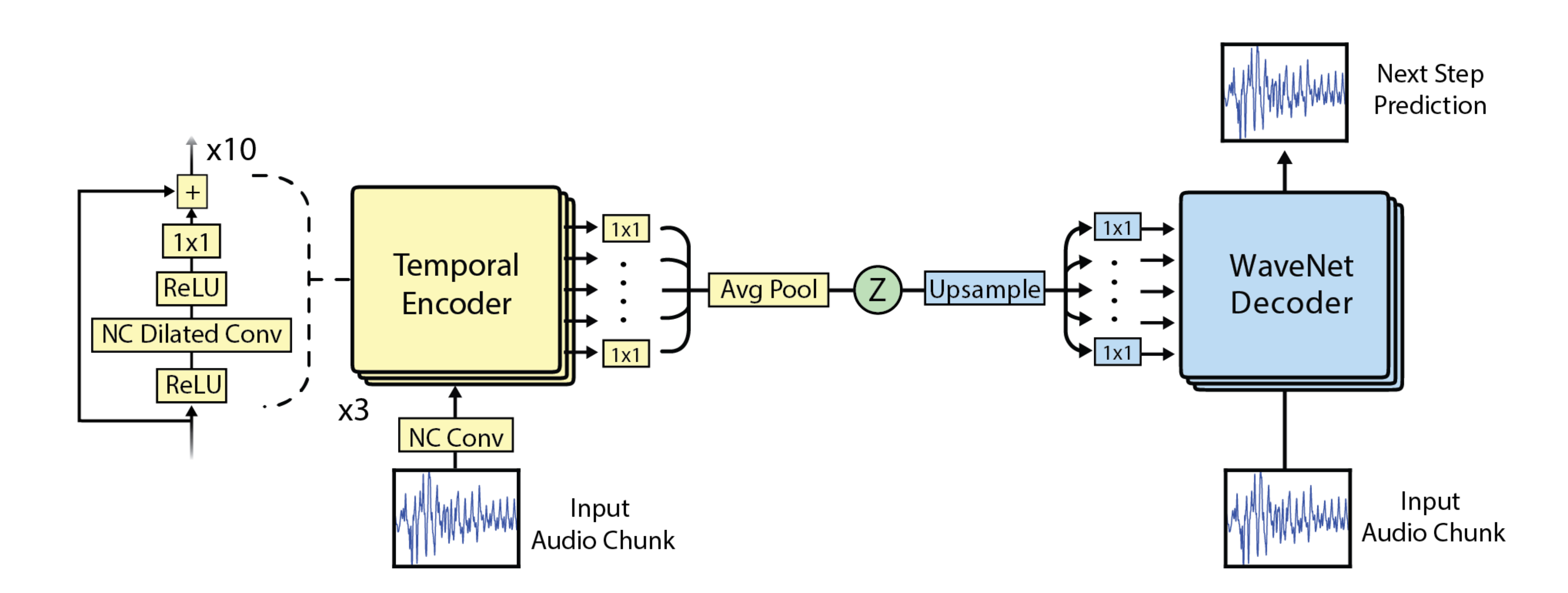
S+N+T Pitch Shifting (timing unchanged)

- Pitch-scale factors [0.89, 0.94, 1.00, 1.06, 1.12]

Spectrogram Synthesis (2017)

NSynth: Neural Audio Synthesis (2017)

- NSynth uses deep neural networks to generate sounds at the level of individual samples
- Audio Morphing in "neural latent space"
- Google Magenta project: `https://magenta.tensorflow.org/nsynth`

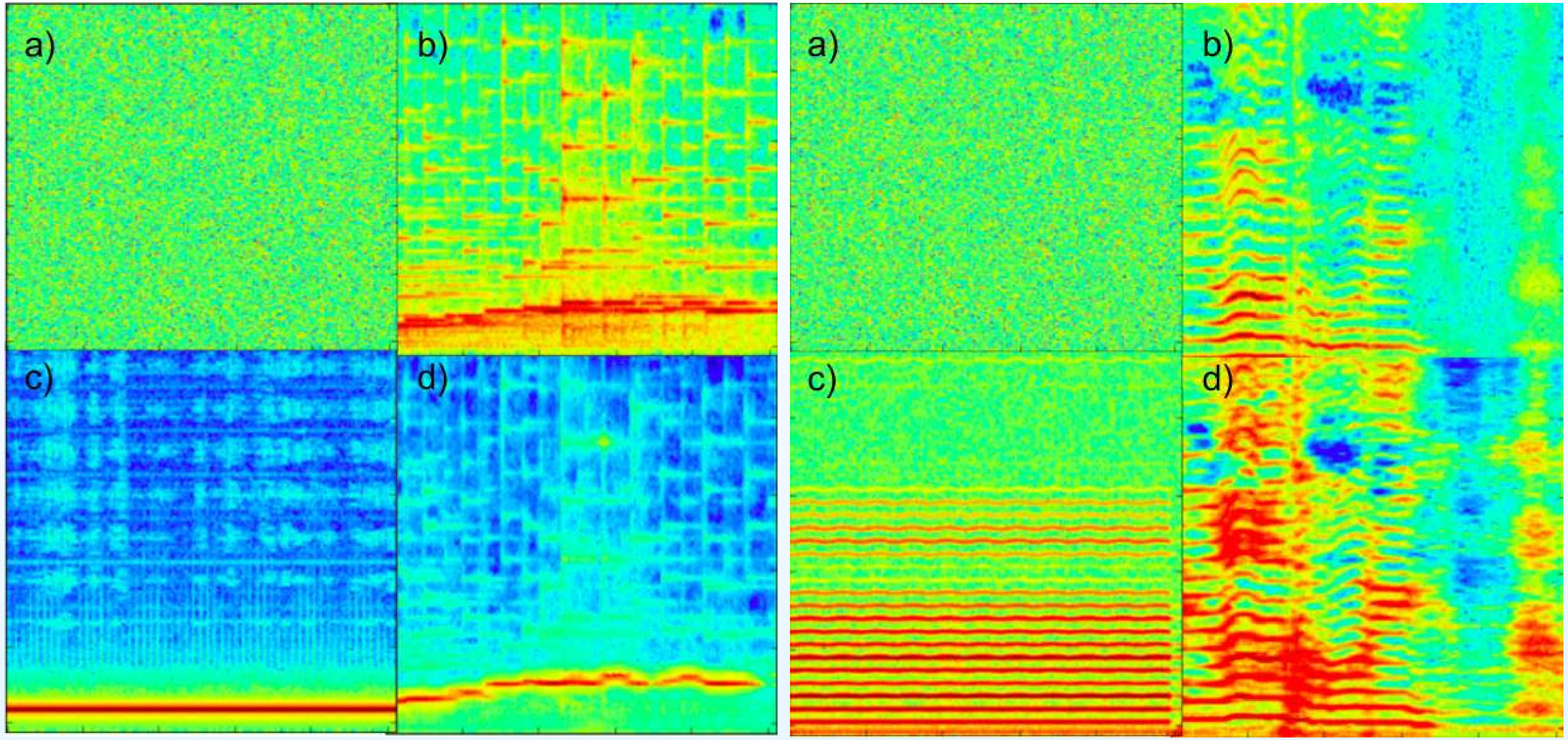


Neural Style Transfer for Audio Spectrograms (2017)

Learned Spectrogram $X(\omega, t)$ minimizes a *sum* of *loss terms:*

- *content loss* = L2 distance between current activation filters and those of "content" spectrogram
- *style loss* = normalized L2 distance between Gram matrix of filter activations of selected convolutional layers chosen as corresponding to "style"
- *differences in temporal and frequency energy envelopes*
- NIPS paper by Verma and Smith (2017): `https://arxiv.org/abs/1801.01589`
- Original paper for images (2015): Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "A neural algorithm of artistic style." arXiv preprint arXiv:1508.06576(2015): `https://arxiv.org/abs/1508.06576`
- Nice intro (2016): `http://yeephycho.github.io/2016/09/14/neural-style/`

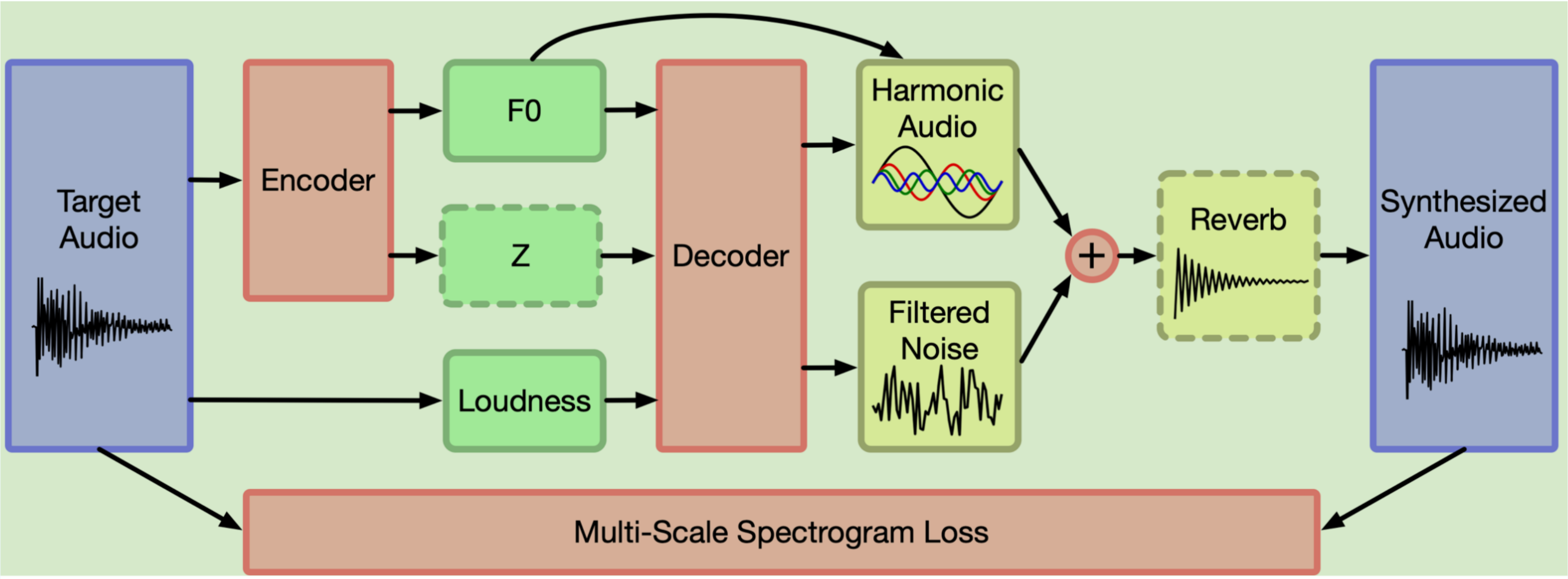
Spectrogram Style Transfer, Continued



- Left: Tuning-fork “style” imposed on a harp sample: Adaptive filtering down to fundamental observed
- Right: Violin “style” imposed on singing voice: Bandwidth extension observed

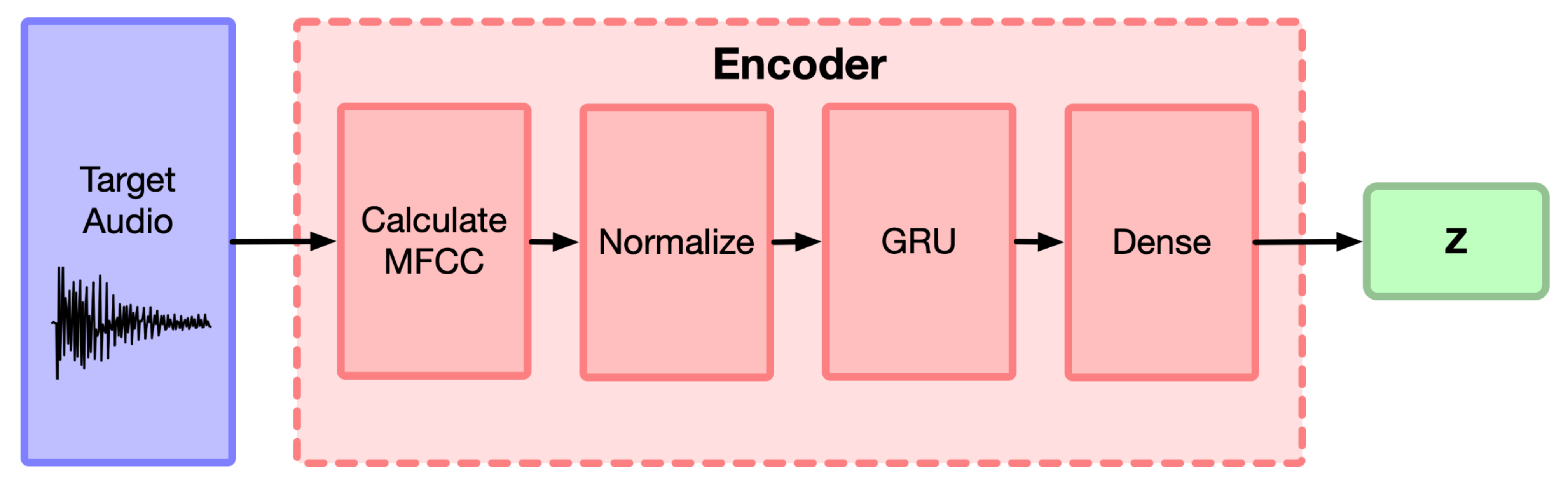
Differentiable DSP (2019)

Differentiable DSP



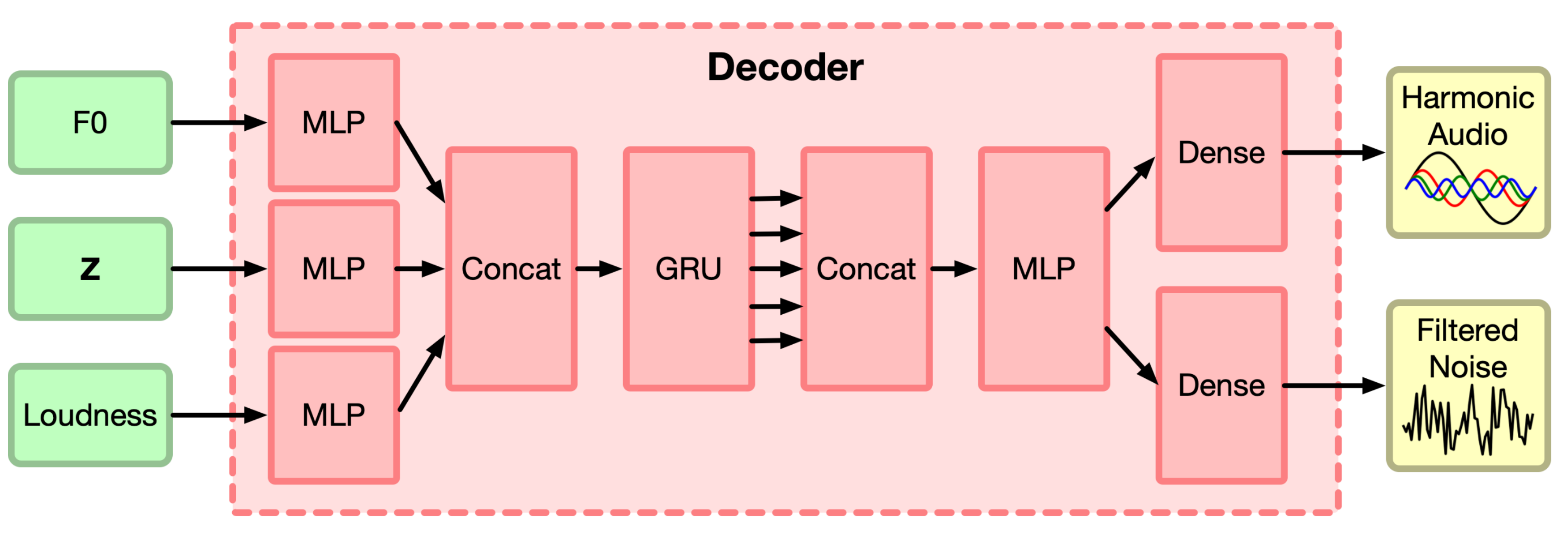
- Jesse Engel et al. at Google Magenta Group
- Neural network analysis/synthesis for *differentiable signal models*
- *Additive Synthesis* example:
  - Loudness normalized by A-weighted log-power spectrum
  - Fundamental Frequency F0 from pretrained CREPE pitch detector
  - Timbre vector Z from *autoencoder*
  - Timbre vector decodes to sinusoidal amplitude trajectories

DDSP Encoder



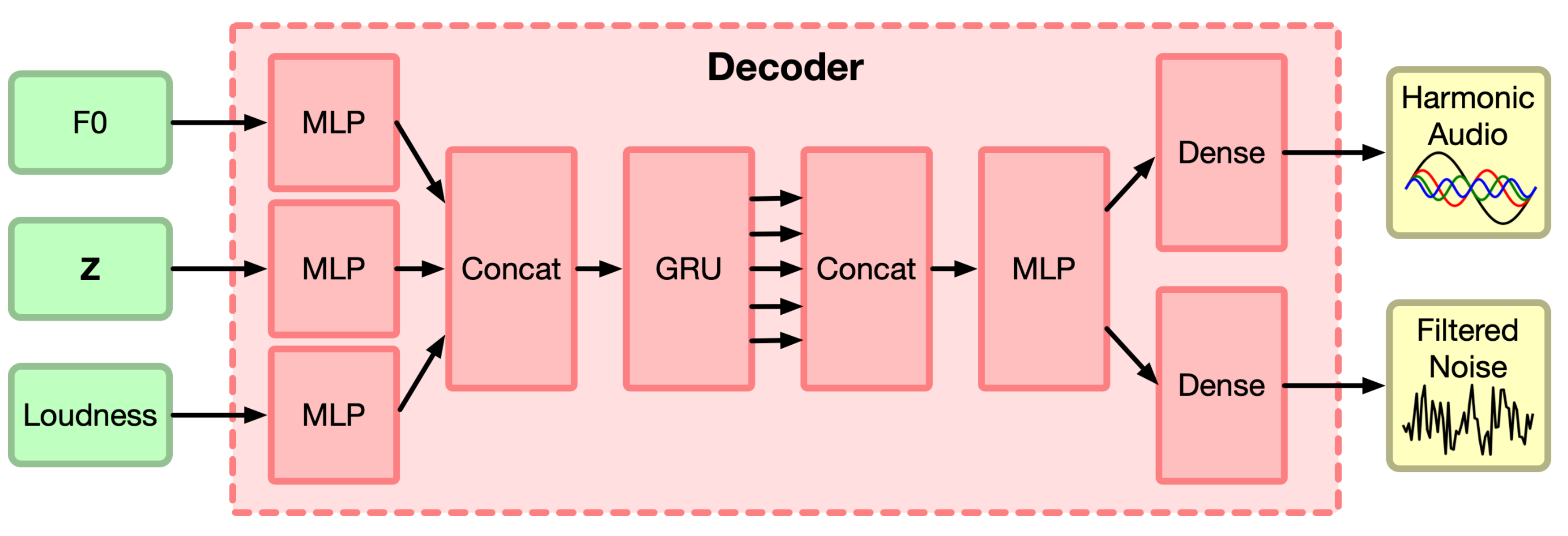
- Loudness and F0 of Target Audio have been normalized away
- MFCC = Mel Frequency Cepstral Coefficients
- GRU = Gated Recurrent Unit (Cho 2014) - similar to LSTM = Long/Short-Term Memory
- Dense = Fully Connected Linear Deep Neural Net (512-to-16 compression step)
- F0 and Loudness normalization leave only *timbre* to be encoded

DDSP Decoder



- MLP = Multi-Layer Perceptron (classical neural network)
- 250 time steps (frames) included
- Output is additive synthesis parameters (sines + filtered noise)

DDSP MLP



- RELU = Rectified Linear Unit (half-wave rectifier)
- 3 layers and 512 Units
- Entire model is differentiable end to end, so back-propagation can optimize everything together (ADAM optimizer used)
- Optimization is generally Stochastic Gradient Descent

Summary and Future Prospects

Spectral Modeling History Highlights

- Bernoulli's modal sums (1733)
- Fourier's initial theorem (1822)
- Telharmonium (1906)
- Hammond organ (1930s)
- Channel Vocoder (1939)
- Phase Vocoder (1966)
- "Additive Synthesis" (1969)
- FFT Phase Vocoder (1976)
- Sinusoidal Modeling (1977,1979,1985)
- Sines+Noise (1989)
- Sines+Transients (1989)
- TF Reassignment (1995)
- Sines+Noise+Transients (1998)

***Perceptual audio coding*:**

- Princen-Bradley filterbank (1986)
- K. Brandenburg thesis (1989)
- *Auditory masking* usage
- Dolby AC2
- Musicam
- ASPEC
- MPEG-I,II,IV (S+N+T "parametric sounds")

***Neural Models (Incomplete!)*:**

- WaveNet (2016)
- SampleRNN (2017) ...

Neural Audio Generation in Recent Years

- WaveNet (2016) [expensive but amazing quality]
- SampleRNN (2017)
  Check out https://dadabots.com/
  e.g., “lofi classic metal ai radio”:
  https://www.youtube.com/watch?v=J1NV6CUJI18
- DDSP (2020)
- JukeBox (2020) [expensive - inspired many offshoots]
- SoundStream (2021) [Multilevel VQ - excellent elementary background]
- PerceiverAR (2022) [SoundStream tokens $\rightarrow$ Perceiver]
- AudioLM (2022) [parallel *semantic* and *acoustic* token sequences]
- Riffusion (2022) [diffusion encoding of spectrograms]
- MusicLM (2023) [based on SoundStream, AudioLM]
- Numerous follow-on papers
  (Track citations in *Google Scholar*)

Future Prospects

Observations:

- Sinusoidal modeling of sound is “Unreasonably Effective”
- Basic “auditory masking” discards $\approx$ 90% information
- Interesting neuroscience observation:

  > “... most neurons in the primary auditory cortex A1 are silent most of the time ...”

  (from “Sparse Time-Frequency Representations”, Gardner and Magnesco, PNAS:103(16), April 2006)
- In addition to evolving our *brains*, we evolved our *inner ear*
  (real-time spectrum analysis hardware)
  - *Efficient* audio modeling focuses on the *spectral peak behavior*
  - We evolved *neural processing* of a *fixed time-frequency analysis*
  - Neural nets can identify, track, and predict “auditory objects”
  - MIT Researchers have replicated auditory cortex activity in neural nets